Combinaison de sources de données pour l'amélioration de la prédiction en apprentissage : une application à la prédiction de la perte de poids chez l'obèse à partir de données transcriptomiques et cliniques

Mohamed Ramzi Temanni

THESE DE DOCTORAT DE
L'UNIVERSITE PIERRE ET MARIE CURIE

Spécialité : Informatique Biomédicale
Ecole Doctorale Santé Publique : Epidémiologie et Sciences
de l'Information Biomédicale

Présentée par

M. Mohamed-Ramzi TEMANNI

Pour obtenir le grade de

DOCTEUR de l'UNIVERSITÉ PIERRE ET MARIE CURIE

Sujet de la thèse :

Combinaison de sources de données pour l'amélioration de la prédiction en apprentissage : Une application à la prédiction de la perte de poids chez l'obèse à partir de données transcriptomiques et cliniques.

Soutenue le 23.06.2009

Devant le jury composé de :
Mme. Karine Clément, Professeur, Université Paris 6, INSERM Co-directrice de thèse
M. Pierre Le Beux, Professeur, Université Rennes 1, Rapporteur
M. Bertrand Guidet, Professeur, Université Paris 6, Examinateur
Mme. Céline Rouveirol, Professeur, Université Paris 13, Rapporteur
M. Jean-François Zagury, Professeur, Chaire de Bioinformatique du CNAM, Examinateur
M. Jean-Daniel Zucker, DR IRD, Université Paris 6, Directeur de thèse

Université Pierre & Marie Curie - Paris 6
Bureau d'accueil, inscription des doctorants et base de données
Esc G, 2ème étage
15 rue de l'école de médecine
75270-PARIS CEDEX 06

Tél. Secrétariat : 01 42 34 68 35
Fax : 01 42 34 68 40
Tél. pour les étudiants de A à EL : 01 42 34 69 54
Tél. pour les étudiants de EM à ME : 01 42 34 68 41
Tél. pour les étudiants de MF à Z : 01 42 34 68 51
E-mail : scolarite.doctorat@upmc.fr

À la mémoire de mon père,

À ma mère, à ma sœur et à mon frère

Remerciements

Au terme de cette thèse, j'aimerais remercier les personnes qui ont contribué à ce travail de recherche, par leur confiance, leur patience, leur compétence et leurs conseils précieux. Mes premiers remerciements iront tout naturellement à mon directeur de thèse, Mr Jean-Daniel Zucker pour l'intérêt qu'il a porté à mes travaux depuis qu'il m'a introduit dans le monde de la bioinformatique. J'aimerais lui témoigner ici toute ma reconnaissance pour ses encouragements ainsi que pour la confiance permanente qu'il m'a accordée. Cela a été pour moi un véritable plaisir de travailler avec lui.

Je suis très reconnaissant à Mr Alain Venot de m'avoir accueilli dans son laboratoire (LIM&BIO) et qui m'a donné l'opportunité de travailler dans la meilleure des conditions.

Je tiens par ailleurs à remercier Mme Karine Clement de m'avoir accueilli dans son équipe (Nutriomique) soudée et amicale et de m'avoir donné la chance de participer à des projets de grande envergure.

Je remercie Mr Pierre Le Beux et Mme Céline Rouveirol d'avoir accepté d'être rapporteurs de mes travaux de recherche. Je tiens également à remercier, Mr Bertrand Guidet, Mr Jean-François Zagury pour avoir voulu participer au jury de thèse.

Je souhaite adresser mes sincères remerciements à Mr Alain-Jacque Valleron directeur de l'école doctorale SPESIB pour son soutien au doctorant sans oublier Mme Evelyne Guilloux qui nous accueille toujours avec le sourire et nous facilite les démarches administratives.

Je voudrais également remercier tous les membres du laboratoire LIM&BIO et en particulier: Anis, Bernard, Blaise, Catherine, Lina, Massoud, Jean-Baptiste, Sajjad, Sylvie et Vahid. Je tiens aussi à remercier les membres de l'équipe 7 "Nutriomique" pour l'ambiance familiale instaurée dans le laboratoire (Christine P, Christine R, Daniel, David, Elise, Flavien, Froogh, Guillaume, Mayoura, Michèle, Nadia, Nadine, Johan, Véro, et Salwa). Un grand merci à Corneliu et Edi pour leurs présences à mes côtés, leurs écoutes et leurs amitiés. Je remercie vivement Mr Younes Bennani et Mr Mustapha Lebbah du LIPN pour leur écoute et leur aide.

Je tiens aussi à remercier Mme Grace Shieh et Mr Gert Lanckriet de m'avoir accueilli dans leur laboratoire et qui m'ont permis d'élargir mes connaissances scientifiques, mais aussi culturelles.

Un grand merci à mes amis, et plus particulièrement Abdelali, Ahmed, Chokri, Eya, Gwen, Hatem, Imene, Karim, Malek, Mehdi, Mohamed, Nabil, Tarek, Timo et Seraya pour leur présence à mes côtés et leurs encouragements.

Ma plus grande reconnaissance va à ma famille pour leurs soutiens et encouragements : ma mère Nabiha, ma sœur Rim, mon frère Mohamed et tous mes oncles et tantes, cousins et cousines un peu partout dans le globe.

A ceux qui, chacun à leur façon, ont fait ce bout de chemin à mes côtés, merci.

Résumé

Les maladies complexes comme l'obésité sont des maladies multifactorielles. Peu de travaux existent pour essayer de prédire les effets des différents traitements et ainsi mieux adapter les traitements aux patients. L'utilisation de modèles prédictifs pour mieux guider le choix des traitements de l'obésité reste un champ de recherche peu exploré malgré le fort impact qu'elle pourrait avoir vu la prévalence de cette maladie. Dans d'autres domaines de la médecine, comme la cancérologie par exemple, de telles méthodes sont déjà utilisées pour l'aide au diagnostic se basant notamment sur des données issues de puces à ADN. Cette technologie s'avère adaptée et son utilisation a donné lieu à des résultats intéressants pour dépister les maladies ou aider les médecins dans leur choix thérapeutique. Cependant si celle-ci s'avère suffisante pour prédire d'une manière satisfaisante dans le domaine du cancer, en revanche elle s'avère d'un apport limité dans le cadre d'une application aux données de l'obésité. Cela suggère l'utilisation d'autres données patients pour améliorer les performances en prédiction. Les travaux de recherche présentés dans ce mémoire abordent les problèmes de la prédiction de la perte de poids suite à un régime ou une chirurgie bariatrique. Nous avons analysé le problème de la prédiction de la perte de poids à partir des données transcriptomique dans le cadre de deux projets européens et aussi à partir des données biocliniques dans le cadre de la chirurgie de l'obésité. Nous avons ensuite proposé trois concepts de combinaisons de modèles : combinaison de données, combinaison de méthodes et combinaison avec abstention. Nous avons analysé empiriquement ces trois approches et les expérimentations ont montré une amélioration des résultats pour les données de l'obésité même si ceux-ci restent bien en deça de ce qu'on observe avec les données cancers.

Table des matières

Liste des tableaux

Liste des figures

Introduction

Les récentes évolutions dans la recherche en génétique, de l'informatique et des technologies de l'information ont entraîné le développement de la bioinformatique dans le domaine de la médecine et de la biologie. La bioinformatique est un domaine de recherche pluridisciplinaire qui allie à la fois la biologie et d'autres disciplines comme l'informatique, les statistiques ou les mathématiques. Cette nouvelle science fait appel aux outils informatiques pour récupérer, traiter et analyser des données biologiques.Les biopuces représentent une des nouvelles technologies émergentes de la recherche biomédicale. Elles permettent de mesurer et de visualiser très rapidement les différences d'expression entre les gènes et ceci à l'échelle d'un génome complet. L'application la plus répandue des biopuces concerne actuellement l'étude du transcriptome dans le cadre d'études différentielles, d'étude de profil génique ou d'études pronostiques ou prédictives. La bioinformatique ne se restreint pas au seul déploiement des outils informatiques, elle se focalise aussi sur l'étude d'une problématique biologique et la recherche de méthodes qui répondentaux questions biologiques en tenant compte du contexte et des spécificités du domaine. La réussite d'une analyse bioinformatique nécessite une bonne compréhension du problème, et cela ne serait pas possible sans une étroite collaboration entre les informaticiens et les biologistes. Plusieurs travaux de recherche en bioinformatique, basés sur les modèles d'apprentissage issus de l'intelligence artificielle et utilisant les données transcriptomiques, ont permis de mieux comprendre certaines maladies complexes comme le cancer.De telles approches sont rares dans le domaine de l'obésité.

L'obésité est une maladie d'origine multifactorielle chronique qui dépend de l'environnement (social et culturel), des facteurs génétiques, physiologiques, métaboliques, comportementaux et psychologiques. Les traitements actuels ne se focalisent plus sur la seule perte de poids, mais sur une approche thérapeutique multidisciplinaire (nutritionniste, endocrinologue, psychiatre, etc.) avec une prise en charge des complications associées. L'obésité favorise un très grand nombre de pathologies qui lui sont donc souvent associées (on parle de comorbidités). La complexité biologique de l'obésité suggère que l'utilisation des puces

à ADN pour identifier des gènes capables d'élucider le fonctionnement du métabolisme du tissu adipeux et son altération au cours de l'obésité peut paraitre une approche prometteuse. Ceci est principalement dû au fait qu'il existe une grande variabilité interindividuelle quant à la réponse à une intervention diététique, variabilité dont l'origine demeure principalement inconnue.

Dans le cadre de cette thèse nous nous sommes intéressés dans un premier temps à l'étude et l'analyse de la prédiction de la perte de poids suite à un régime à partir des données transcriptomiques et cela dans le cadre de deux projets européens (Nugenob et Diogenes). Cette étude consiste à mettre en place des modèles de prédiction et à trouver des listes de gènes prédicteurs qu'il faudrait explorer en regardant les informations qui leurs sont associées dans les bases de données publiques et aussi leurs associations avec les travaux déjà réalisés dans notre équipe. Nous avons aussi analysé le problème de la prédiction de la perte de poids suite à une intervention chirurgicale à partir des données biocliniques. Du fait de l'importance de chacune des sources de données utilisées dans nos analyses,notre attention s'est portée sur la combinaison des sources et cela afin d'améliorer la prédiction de la perte de poids en appliquant deux approches : la combinaison de données et la combinaison de modèles.

Structure du mémoire

Chapitre 1 : Introduction à l'épidémiologie de l'obésité:
vers l'intégration de données hétérogènes

Dans ce chapitre, nous parlons du cadre général dans lequel s'inscrit cette thèse, celui de l'épidémiologie de l'obésité. Dans la première partie de ce chapitre, nous introduisons brièvement les principes de l'épidémiologie humaine, ses concepts et ses différentes approches en passant par l'épidémiologie génétique et l'émergence d'un nouveau panel d'outils afin de répondre aux exigences de ce domaine.

La deuxième partie de ce chapitre introduit le problème de l'obésité à l'échelle mondiale, reporte les chiffres alarmants de l'obésité dans le monde, en Europe et en France et les conséquences de cette maladie sur les individus et les organismes de santé publique.

Dans la dernière partie de ce chapitre, nous parlons des efforts mis en place à différentes échelles pour la prévention et la prise en charge de cette maladie. Nous mettons l'accent sur l'importance des données recueillies dans les différentes études cliniques et soulignons l'apport de ces données pour la compréhension et la mise en place du système d'aide à la décision médicale qui pourrait faciliter la prise en charge des patients et améliorer ainsi leur vie au quotidien.

Chapitre 2 : Des données biologiques aux données transcriptomiques

Après une introduction biologique de l'organisme et ses constituants, nous reprenons les principes de la biologie à haut-débit et de la science des 'omiques'. Ensuite, nous expliquons le design ainsi que la mise en place d'une expérimentation de puce à ADN. Par la suite, nous parlons de l'exploitation des puces ADN dans les études cliniques et biologiques. Enfin, nous

détaillons les données utilisées dans le cadre des différents travaux de rechercheréalisés dans le cadre de cette thèse.

Chapitre3 : Aspects méthodologiques de la fouille de données biomédicales

Dans ce chapitre nous parlons des différentes méthodes de fouille de données utilisées dans le domaine biomédical et plus précisément celles adaptées à l'analyse des données transcriptomiques. Nous parlons brièvement du principe des approches de l'apprentissage non-supervisé tel que les cartes auto-organisatrices, la classification hiérarchique et la classification par nuées dynamiques en citant les travaux phares liés à ces approches. Par la suite, nous évoquons les approches supervisées que nous avons expérimentées dans le cadre de nos travaux de recherche et leur application en transcriptomique. Nous finissons ce chapitre par l'explication des méthodes d'estimation des performances des modèles en apprentissage supervisé.

Chapitre 4 : Prédiction de la perte de poids chez les patients obèses

Dans ce chapitre, nous détaillons trois études relatives à la prédiction de la perte de poids. Deux d'entre elles ont été réalisées dans le cadre de deux projets européens (Nugenob et Diogenes) et ont pour but de déterminer des modèles de prédiction de la perte de poids après un régime faible en calorie à partir de données transcriptomiques. La troisième étude quant à elle s'intéresse à la prédiction de la perte de poids après une intervention chirurgicale de type bypass à partir des données biocliniques. Pour cette analyse, nous nous sommes aussi intéressés à la prédiction de paramètres biocliniques autre que le poids et qui peuvent être des indicateurs de l'amélioration de l'état de santé après une intervention chirurgicale. Nous avons aussi investigué l'étude de profils de variation des paramètres biocliniques des patients au cours du temps et la prédiction de cette évolution.

Chapitre 5 : Améliorer la prédiction à partir de la combinaison de données cliniques et transcriptomiques

Dans ce chapitre, nous introduisons l'état de l'art de la combinaison de données et présentons les différentes stratégies de combinaison déployées dans des travaux de recherche dans le domaine biomédical. Ensuite, nous présentons deux modèles d'apprentissage basés sur la combinaison des données cliniques et transcriptomiques pour améliorer la prédiction de la perte de poids chez des sujets obèses. Le premier est un moyen de combiner les données et d'évaluer la pertinence de cette combinaison. Le deuxième est basé sur une approche d'abstention/délégation. Toute deux sont des approches exploratoires mises en place pour améliorer l'apprentissage à partir de deux sources.

Conclusion : Dans ce dernier chapitre, nous faisons une synthèse de travaux de recherche que nous avons réalisés au cours de cette thèse et évoquons les améliorations qui peuvent être apportées à ce travail dans le futur.

Chapitre 1

Introduction à l'épidémiologie de l'obésité: vers l'intégration de données hétérogènes

1.1 Épidémiologie humaine

Tout être humain est unique non seulement par son aspect physique mais aussi par son caractère, sa personnalité et ses comportements. L'historique de la santé de chaque personne est de la même manière unique aussi. De fait, certaines personnes sont sensibles aux infections virales telle que la grippe par exemple tandis que d'autres sont plus résistantes. Ce caractère n'est pas facilement prédictible pour chaque individu, néanmoins il est évident que les caractéristiques et le comportement de chacun jouent un rôle dans la causalité et le degré d'affection aux maladies. Si une population d'individus est exposée d'une manière identique à la même cause d'une maladie on s'attend à ce que ces individus soient atteints plus ou moins de la même manière par cette maladie. De ce fait et afin de comprendre l'incidence d'une maladie sur un individu, l'étude de l'effet de cette maladie sur une population bien définie est nécessaire. L'enjeu est la meilleure compréhension de la maladie et une prise en charge plus adaptée aux patients.

1.1.1 **Définition de l'épidémiologie**

L'*épidémiologie* est l'étude de la répartition et des déterminants des maladies dans les populations.Le terme provient du mot 'epidemic' qui lui-même viendrait du mot 'epidemeion', un mot employé par Hippocrate pour décrire une maladie qui 'visitait les gens'. Elle vise à la compréhension des causes des maladies et à l'amélioration de leurs traitements et des moyens de les prévenir.

L'ensemble des champs couverts par la santé publique repose sur les données épidémiologiques. Il est important de comprendre que les études épidémiologiques comparent les individus atteints aux individus sains sous forme de groupes ou de populations. Il peut en être de même pour l'influence d'un risque auquel une population est exposée qui sera mis en évidence par rapport à une population non exposée (témoin).

La ***distribution*** de la maladie étudiée est généralement géographique mais des distributions par âge, sexe, classe sociale, ethnicité sont toujours d'intérêt. Parfois, la même population géographique est comparée à elle-même à différents temps pour explorer l'évolution d'une maladie.Les ***déterminants*** d'une maladie sont les facteurs provocateurs de la maladie. L'étude de la distribution de la maladie est essentiellement descriptive. L'étude des déterminantsquant à elle,vise l'étiologie de la maladie.

L'objectif de l'épidémiologie est d'informer les professionnels de la santé et la population plus généralement, des améliorations de santé qui peuvent être faites par l'intermédiaire des approches descriptives et étiologiques. Les analyses descriptives doivent permettre une meilleure allocation des services de santé. Les analyses étiologiques doivent permettre d'agir sur les causes et réduire les chances de développer telle ou telle maladie. Les données épidémiologiques sont des sources importantes pour la planification et l'évaluation des services de santé.

L'épidémiologie est souvent vue comme une branche de la médecine qui s'intéresse à la population plutôt qu'aux individus. Alors que les praticiens hospitaliers s'intéressent à trouver le meilleur conseil ou traitement à donner à chaque patient individuellement, les épidémiologistes s'intéressent plutôt à trouver un conseil destiné à une population afin de réduire l'effet et l'étendu de la maladie. Cependant, comme l'épidémiologie utilise des données d'agrégation de personnes, elle est considérée comme étant une branche appliquée de la statistique. Les avancées dans ce domaine ont été réalisées grâce à l'interaction entre les différentes disciplines de la médecine et les statistiques. Parmi les autres disciplines représentées dans les groupes de recherche en épidémiologie, citons les biochimistes, les généticiens, les sociologues et les informaticiens. D'autres professionnels peuvent intervenir

aussi comme les nutritionnistes et les économistes dans le cadre d'études plus ciblées. Une illustration de l'histoire de l'épidémiologie est présentée plus en détail dans le livre de Stolley et Lasky (Stolley and Lasky 1995).

1.1.2 **Études épidémiologiques**

Il existe un nombre croissant de modèles d'étude utilisés en épidémiologie et les étiquettes employées pour les décrire sont nombreuses. Cependant, cinq catégories peuvent être répertoriées :

- **Séries de cas** (Clinique et population): description d'une série de cas comparables, mais sans comparaison avec un groupe témoin ou un autre groupe de cas.

- **Transversale**: description de la fréquence d'une maladie, de ses facteurs de risque ou de ses autres caractéristiques dans une population donnée pendant un laps de temps déterminé. Comparaison des données obtenues en fin d'études à celles du début de l'étude : étude d'une association (et non d'une relation causale) entre une intervention donnée et l'issue clinique

- **Cas-témoin** : étude d'observation rétrospective dans laquelle les caractéristiques des malades (les cas) sont comparées à celles de sujets indemnes de la maladie (les témoins). Particulièrement adaptée pour les maladies rares ou celles qui présentent une longue période entre l'exposition et l'issue et pour l'étude d'hypothèses préliminaires

- **Cohorte** (prospective and rétrospective) : étude d'observation, le plus souvent prospective, dans laquelle un groupe de sujets exposés (à des facteurs de risque d'une maladie ou à un traitement particulier) est suivi pendant une période déterminée et comparée à un groupe contrôle non exposé. Étude éventuellement rétrospective réalisée sur base des dossiers médicaux, par exemple, pour évaluer les risques auxquels les sujets ont été exposés antérieurement

- **Essai** : étude expérimentale, où les patients éligibles, sélectionnés pour une intervention thérapeutique, sont répartis de manière aléatoire en 2 groupes : le premier groupe reçoit le traitement, tandis que le second reçoit en général un placebo. Répartition au hasard ayant pour but d'assurer que les patients répartis dans les 2 groupes de l'essai sont rigoureusement semblables en tous points, excepté en ce qui concerne l'intervention projetée. Réalisation de l'étude en aveugle ou en double aveugle de manière à écarter tout biais éventuel.

La confusion entre ces cinq modèles est classique. Elle est accentuée par l'utilisation des différents termes et l'apparition continuelle de nouveaux mots. Afin de mieux comprendre les différents types d'études épidémiologiques, reprenons dans la Table 1 les idées essentielles ainsi que les objectifs de recherches pour chaque type d'étude.

Type d'études	Idées essentielles	Quelques objectifs de recherches
Série de cas	• Compter des cas et liés aux données d'une population • Observer les caractéristiques des cas pour mettre en évidence des hypothèses causales	• Etudier les signes et les symptômes et créer une définition de la maladie • Surveiller la mortalité/le taux de morbidité • Trouver des associations • Générer/tester des hypothèses
Transversale	• Étudie l'état de santé et des maladies dans une (des) population(s) dans un cadre espace-temps défini	• Mesurer la prévalence d'une maladie ou de ses facteurs. • Trouver des associations entre les maladies et les facteurs apparentés
Cas-témoins	• Analyser les différences et les similarités entre deux séries	• Trouver des associations • Générer/tester des hypothèses

Cohorte	• Suivre les informations sur les facteurs de risques et l'état de santé d'une population particulière	• Etudier l'histoire naturelle de la maladie. • Mesurer l'incidence de la maladie • Lier les maladies aux causes probables/ Trouver des associations • Générer / tester des hypothèses
Essai	• Retrouver des mesures pour améliorer la santé ensuite suivre des groupes afin de quantifier cette amélioration.	• Tester la compréhension des causes • Étudier comment influencer l'histoire naturelle de la maladie • Evaluer les bienfaits et le coût de l'intervention

Table 1: Types d'études épidémiologiques et leur application [d'après (Bhopal 2002)]

Dans le cas où une étude est atypique, ou comprend un mélange d'idées, il est important de comprendre les idées sous-jacentes à la conception de l'étude épidémiologique, notamment en termes de finalité, de forme, d'analyse, d'interprétation, et la base de la notion de population. Cette compréhension permet de définir les points communs de la conception des études épidémiologiques, l'objectif commun de ces études étant la compréhension de la fréquence, du motif, et des causes de la maladie dans les populations. Ces études épidémiologiques sont toutes ancrées dans le concept de la population, la connaissance de la relation entre la population étudiée et la population source est essentielle pour l'interprétation, la généralisation et la compréhension des données. Avant d'entamer un plan d'étude épidémiologique, il est important de se poser les questions suivantes : où et quand l'étude a été faite ? À quelle population appartient le groupe étudié ? Quelles sont les caractéristiques de l'étude ? Est-ce que les conclusions sont généralisables à toute la population locale et est-ce qu'elles restent vraies dans d'autres populations ?

Type d'analyse	Descriptive/ Analytique	Rétrospective/ Prospective	Observationnelle/ Expérimentale	Débuté avec maladie/ causes de la maladie	Groupe spécifique de comparaison/
Série de cas	Descriptive	Rétrospective	Observationnelle	Maladie	Non
Transversale	Descriptive	Rétrospective	Observationnelle	Maladie / Causes	Non (en général)
Cas-témoins	Analytique	Rétrospective	Observationnelle	Maladie	Oui
Cohorte	Analytique	Rétrospective & Prospective	Observationnelle	Causes (en général)	Oui (en général)
Essai	Analytique	Prospective	Expérimentale	Malaadies (en genral) Cause(parfois)	Oui (avec des exceptions)

Table 2: Classification dichotomique des études épidémiologiques [d'après (Bhopal 2002)]

La population de base est le point de départ de toute expérimentation. Toutes les études épidémiologiques permettent de comparer des maladies dans une ou plusieurs périodes temporelles, personnes ou lieux. Elles contribuent toutes à la mesure de l'impact de la maladie ou des facteurs de risques ainsi que l'étude de la relation entre la maladie et les facteurs à la base de cette maladie. Un des critères importants de la causalité est la consistance, cela demande des hypothèses construites à partir de plus qu'une seule étude et de préférence utilisant des types d'études différents.

Afin de permettre le bon choix du type d'étude épidémiologique, il est essentiel de comprendre le but de l'analyse et cela pour choisir le bon type d'analyse. Il existe plusieurs classifications des types d'analyses qui sont mises en place pour faciliter le choix de la bonne

analyse à suivre pour une étude bien définie. 3 dichotomies sont principalement utilisées pour distinguer les études épidémiologiques : descriptive / analytique; rétrospective/prospective ; observationnelle / expérimentale. L'étude descriptive est celle qui fournit des informations à propos des maladies et des facteurs de risques sans pour autant s'intéresser aux causes. L'étude analytique s'intéresse à l'exploration d'hypothèses liées aux causes de la maladie. Une étude rétrospective s'intéresse aux données collectées dans le passé alors que les études prospectives s'intéressent aux données du futur. L'étude observationnelle est celle où l'investigateur observe la progression naturelle d'un événement de santé. Dans une étude expérimentale, les conditions sont fixées. D'autres alternatives dichotomiques peuvent être utiles, par exemple l'absence ou la présence de la maladie au début de l'étude ou encore l'intégration d'un groupe de référence d'ans l'étude ou non.

Pour une lecture plus approfondie sur les études épidémiologiques le lecteur peut se référer au livre de Bhopal (Bhopal 2002; Bhopal 2008) pour un approfondissement conceptuel dans le sujet et celui de Woodward (Woodward 1999) pour une approche orientée conception et design.

1.1.3 **L'épidémiologie génétique**

Les maladies *complexes* comme le cancer ou l'obésité sont des maladies multifactorielles causées par l'interaction entre des facteurs environnementaux, comportementaux, psychologiques, sociaux, sanitaires et génétiques. Ces déterminants influencent la santé de façon indépendante, mais ils s'influencent aussi les uns les autres. Les facteurs psychologiques et comportementaux influencent le statut social de l'individu et vice versa ; la variation génétique est l'un des nombreux déterminants de la susceptibilité aux facteurs environnementaux.

Historiquement, génétique humaine et épidémiologie ont eu des domaines d'intérêt très différents. L'objet principal de la génétique était l'identification de mutations rares causant des syndromes monogéniques. Les succès de la génétique sont nombreux et incluent la découverte des gènes de la mucoviscidose ou de la chorée de Huntington (Kerem, Rommens et al. 1989;

Tibben, Duivenvoorden et al. 1994). Ces gènes sont utilisés en clinique pour le diagnostic. Les épidémiologistes classiques étudiaient les associations entre des facteurs de risque exogènes et des maladies, assurant que le bagage génétique de chacun ne jouait qu'un faible rôle. Les succès de l'épidémiologie sont aussi importants et incluent les causes du choléra, du béribéri et du cancer du poumon (Fraser 1998; Alberg and Samet 2003; Codeco and Coelho 2006).

Aujourd'hui, la génétique et l'épidémiologie cherchent à élucider ensemble les causes des maladies complexes. Les progrès de la biologie moléculaire et le projet du génome humain (Human Genome Project) ont élargi les outils d'étude les mécanismes pathogénétiques des maladies humaines (Khoury 1997; Vaessen and van Duijn 2001). Sur le plan technologique, nous pouvons à présent caractériser des variations de séquences d'ADN à une large échelle (Ellsworth and Manolio 1999; Ellsworth and Manolio 1999). Dans la même période, les concepts génétiques se sont lentement intégrés aux méthodes épidémiologiques. Ceci se reflète aujourd'hui dans la définition de l'épidémiologie génétique, qui est l'étude du rôle des facteurs génétiques et de leur interaction avec des facteurs environnementaux dans la survenue de maladies au sein des populations humaines (Guyatt, Walter et al. 1987; Plomin, Owen et al. 1994; Risch and Merikangas 1996; Khoury and Yang 1998; Ellsworth and Manolio 1999; Ellsworth and Manolio 1999; Ellsworth and Manolio 1999).

1.2 L'obésité : une épidémie des temps modernes

L'émergence de l'obésité remonte au temps d'Hippocrate il y a 2400 ans et qui avait constaté que la mort subite était plus fréquente chez les « gras » que chez les maigres. Voilà que 600 ans plus tard, Galien donne une vision plus précise en décrivant l'obésité comme source de maladie et en proposant une explication physiopathologique et un traitement : «L'obésité est inutile aux hommes et aux femmes car elle les rend malades. ». Galien indique qu'il ne faut pas confondre la rondeur avec la polysarkia, c'est-à-dire l'obésité massive. Il oppose ainsi le surpoids banal, et l'obésité maladie. Vers 1817, un médecin belge, Adolphe Quételet, va définir un indice de masse corporelle (en anglais BMI : Body Mass Index). L'IMC se calcule par une formule qui exprime la corpulence par le rapport du poids sur la taille au carré (IMC = poids [en kg] / taille^2

[en mètres]). Le poids « normal » est compris entre un IMC de 18,5 et de 24,9 kg/m². Le surpoids est défini par un IMC supérieur ou égal à 25 kg/m², l'obésité par un IMC supérieur ou égal à 30 kg/m².

Etat	IMC
Maigreur	<18.5
Etat de référence	18.5 – 24.9
Surpoids	25 – 29.9
Obésité	30 – 34.9
Obésité sévère	35 – 39.9
Obésité massive	≥ 40

Table 3 : Classification de l'état nutritionnel chez l'adulte en fonction de l'indice de masse corporelle (IMC) selon l'OMS et l'International Obesity Task Force

L'obésité est une maladie d'origine multifactorielle chronique qui dépend de l'environnement (social et culturel), des facteurs génétiques, physiologiques, métaboliques, comportementaux et psychologiques. Les traitements actuels ne se focalisent plus sur la seule perte de poids, mais sur une approche thérapeutique multidisciplinaire (nutritionniste, endocrinologue, psychiatre, etc.) avec une prise en charge des complications associées (Basdevant 2000; Basdevant 2003).

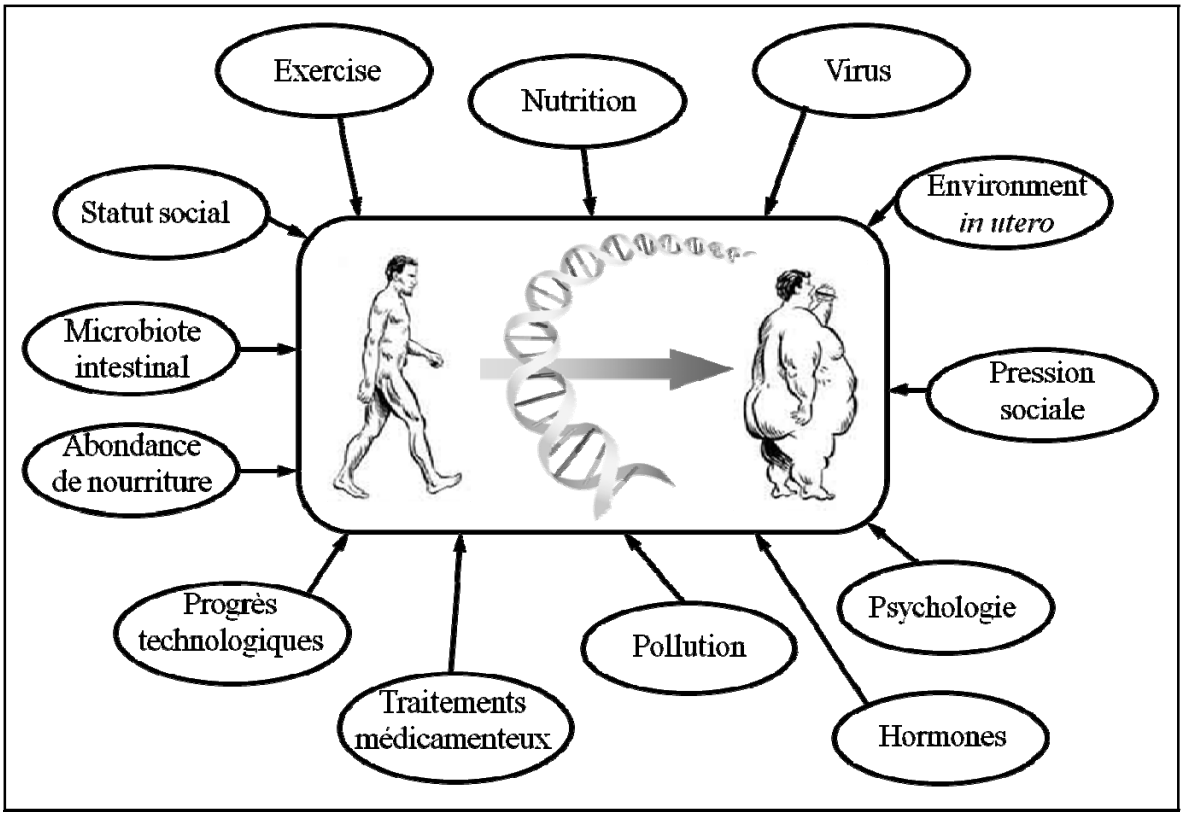


Figure 1: Causes de l'obésité (d'après Mutch & al.)

L'obésité favorise un très grand nombre de pathologies qui lui sont donc souvent associées (on parle de comorbidités). Les principales maladies associées à l'obésité sont : les maladies métaboliques (diabète, dyslipidémies, etc.), les maladies cardiovasculaires, les maladies respiratoires (dyspnées, apnées du sommeil, etc.), des maladies de peau (hypersudation, mycoses), l'arthrose, le reflux gastro-œsophagien, l'incontinence urinaire, certains troubles gynécologiques, la dépression, certains cancers. Il faut lui ajouter les phénomènes de discrimination et de stigmatisation, beaucoup trop fréquents, pouvant conduire certaines personnes à l'isolement social (Bray 1996; Pégorier 2007).

1.2.1 **L'obésité dans le monde**

Cette maladie se développe à un rythme alarmant à travers le monde au cours des deux dernières décennies. Le développement de l'obésité a été déclaré comme étant une épidémie mondiale par l'Organisation mondiale de la santé. De ce fait, un nouveau terme, « *globésité* », a

été attribué pour décrire l'apparition massive du surpoids et de l'obésité à travers la population mondiale. Quelle en est l'ampleur de cette épidémie ? D'après l'Organisation Mondiale de la Santé, plus d'un milliard d'adultes sont en surpoids et au moins 315 millions sont cliniquement obèses. 22 millions d'enfants de moins de 5 ans sont en surpoids ou obèses dans le monde (NCHS 2005; Collins 2006; Collins 2006; WHO 2006).

Ce qui est plus préoccupant est que l'obésité a été liée à un large spectre de maladies dégénératives incluant des désordres métaboliques et certaines formes de cancers. Les sondages ont montré que 80% des diabètes de types 2, 70% des maladies cardiovasculaires et 42% du cancer du sein et du colon sont liés à l'obésité(Collins 2006; Collins 2006). L'obésité est la cause principale pour 30% des cas de dysfonctionnement de la vésicule biliaire donnant lieu à une opération chirurgicale et 26% d'incidence sur hypertension artérielle.

Les conséquences de la globésité ont généré une série de stratégies de perte de poids(Collins 2006; Collins 2006), les produits et les programmes qui induisent à une perte de poids rapide et qui perturbent l'homéostasie métabolique attirent l'attention des centres de marketing et des consommateurs. Cependant, une perte de poids rapide est potentiellement mauvaise pour la santé et induit souvent un gain de poids indésirable par la suite. En outre, de nombreux produits pharmaceutiques anti-obésité sont accompagnés de réactions indésirables, rendant le remède pire que la maladie elle-même. Il est donc très important de développer une stratégie d'intervention thérapeutique en utilisant des suppléments naturels soutenus par une recherche crédible.

1.2.2 **L'obésité en Europe**

L'obésité est relativement courante en Europe, en particulier chez les femmes. Cette maladie est plus marquée dans le sud et les pays d'Europe orientale. On remarque une augmentation des niveaux de surpoids et d'obésité chez les adultes dans toute l'Europe, bien que les taux de prévalence soient différents. Selon un rapport de l'International Obesity Task Force (IOTF) rendu public en mars 2005, la prévalence de l'obésité dans les pays européens varie de 10 à 27% pour les hommes et jusqu'à 38% pour les femmes.

Pays l'UE	année de l'enquête	Hommes (en %)			Femmes (en %)		
		Surpoids*	**obésité****	**total**	**surpoids***	**obésité****	**total**
Royaume-Uni	2006	44,7	24,9	69,5	32,9	25,2	58,0
Allemagne	2003	45,5	20,5	66,0	29,5	21,1	50,6
France	2006	41,0	16,1	57,1	23,8	17,6	41,4
Espagne	2003	46,7	13,9	60,6	30,6	15,1	45,7
Pays-Bas	1998-2002	43,5	10,4	53,9	28,5	10,1	38,6
Italie	2005	42,5	10,5	53,0	26,1	9,1	35,2
Grèce	2003	41,2	26,0	67,1	29,9	18,2	48,1
Malte	2003	46,5	22,9	69,4	34,3	16,9	51,2
Ensemble des 27	-	42,8	16,2	59,0	29,5	18,1	47,5

Table 4 : Prévalence *de la surcharge pondérale, du surpoids et de l'obésité chez les adultes dans l'Union européenne (Source : IOTF) * indice de masse corporel (IMC) compris entre 25 et 29,9** indice de masse corporel (IMC) supérieur à 30*

Si on se restreint uniquement à l'obésité, neuf pays européens au moins ont des taux d'obésité, chez les hommes au-dessus de 20%, y compris la Grèce et Chypre (27%). La prévalence de l'obésité a augmenté d'environ 10 à 40% dans la plupart des pays européens au cours de la dernière décennie. Aux Pays-Bas, l'obésité a augmenté progressivement, passant de 6,2% à 9,3% chez les femmes et de 4,9% à 8,5% chez les hommes depuis la fin des années 1970 jusqu'au milieu des années 1990. La plus forte augmentation a été enregistrée au Royaume-Uni, où le taux d'obésité a augmenté de 13,2% à 22,2% chez les hommes et de 16,4% à 23,0% chez les femmes en 10 ans, jusqu'en 2003.

1.2.3 **L'obésité en France**

L'enquête Obépi (Charles, Eschwege et al. 2008) reporte que depuis 1997, la prévalence de l'obésité est en progression régulière : l'augmentation relative de la population des personnes obèses a été de +9,7% en 3 ans. Cette progression relative avait été de +17% entre

1997 et 2000 et de +17,7% entre 2000 et 2003. Au total, l'augmentation moyenne relative de l'obésité est de +5,7% par an depuis 9 ans. Entre 2003 et 2006, une tendance à l'infléchissement de la courbe de progression de l'obésité est constatée. En extrapolant ces résultats à la population française, on dénombrerait, en 9 ans, 2 347 000 nouveaux cas d'obésité.

L'obésité massive a doublé en 6 ans, progressant de 0,3% à 0,6%. De plus, 42,5% des personnes en surpoids ou obèses présentent au moins un facteur de risque (hypertension artérielle, excès de cholestérol, de lipides ou de triglycérides, diabète) alors que seuls 18,6% des personnes de poids normal en présente un.

Pour combattre cette épidémie, le traitement idéal doit permettre de diminuer le poids à long terme et le maintenir, être moins dangereux que l'histoire naturelle de la maladie et être supportable pour le patient. Ce traitement, au mieux, conduit par une équipe multidisciplinaire, fait appel à une diminution des apports énergétiques et la mise en place d'un équilibre alimentaire et à l'activité physique. L'usage des médicaments est limité en termes d'efficacité et peut conduire à des complications. Quant à la chirurgie gastrique de l'obésité, elle reste un traitement de seconde ligne, réservée aux obésités massives ou sévères avec comorbidités. Les sujets porteurs d'une obésité morbide sont fragiles, surtout s'il existe des maladies associées, c'est pourquoi une évaluation multidisciplinaire préopératoire précise est nécessaire pour prévenir ces risques.

Le régime à très faibles calories VLCD[1](Saris 2001; Clement, Viguerie et al. 2004) sur une courte période (<6 mois) est une alternative pour perdre rapidement du poids et améliorer les comorbidités. Bien que dans ce type de régime les apports alimentaires soient strictement contrôlés (hospitalisation, surveillance médicale), il existe une variabilité inter individuelle quant à la perte de poids obtenue à l'issue de ce régime.

[1] Very Low Calorie Diet

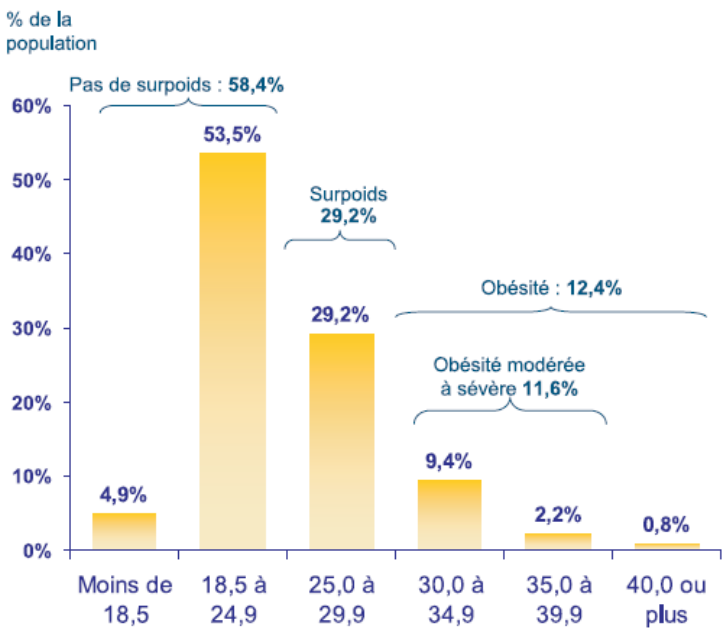


Figure 2 : Répartition des niveaux d'I.M.C en France en 2006 (d'après l'enquête obépi-roche)

Les déterminants de cette perte de poids ne sont pas clairement définis. Par ailleurs, les laboratoires de nutrition disposent désormais de ressources diverses (biothèques, sérothèques, banques de tissus, etc.). Parmi ces ressources, le tissu adipeux est une ressource importante pour comprendre les mécanismes de l'obésité.

1.2.4 **Le tissu adipeux : rôle central dans l'homéostasie énergétique.**

Le tissu adipeux a été longtemps considéré comme étant un tissu «passif» dévolu au stockage et à la mobilisation des lipides en réponse à des signaux hormonaux ou neuronaux. La présence de très nombreux récepteurs à l'insuline, à l'hormone adrénocorticotrope ou à l'adrénaline est en accord avec ce concept. Toutefois, la découverte de la leptine en 1994(Zhang, Proenca et al. 1994) et de l'adiponectine en 1995 (Scherer, Williams et al. 1995) a initié la reconnaissance de ce tissu comme un organe sécrétagogue à part entière.

La leptine est une hormone de 16 kDa encodée par le gène ob (obèse) et sécrétée quasi exclusivement par l'adipocyte. La leptine agit sur son principal tissu cible, l'hypothalamus après s'être liée à un récepteur de type « cytokine » qui lui est propre, ob-Rb (Guerre-Millo 2002). La leptine agit comme un lipostat. En effet, en réponse à l'augmentation des réserves adipeuses, elle va provoquer l'arrêt de la prise alimentaire et exercer ainsi un rétrocontrôle négatif sur la masse adipeuse. À l'inverse, une diminution des réserves adipeuses va réduire la sécrétion de leptine entraînant une reprise alimentaire. Chez l'homme, il existe une relation positive entre leptinémie et masse grasse corporelle. Les taux plasmatiques de leptine avoisinent ~ 5 ng/ml chez le sujet normal et sont 10 fois plus élevés, soit ~ 50 ng/ml, chez le sujet obèse.

L'adiponectine est une protéine de 30 kDa également sécrétée essentiellement par les adipocytes. L'adiponectine agit sur ses principaux tissus cibles (muscles, foie, cellules endothéliales, cellules hématopoïétiques) après s'être liée à des récepteurs spécifiques appartenant à une nouvelle famille de récepteurs (Goldfine and Kahn 2003). Ceux-ci comprennent sept domaines transmembranaires, mais sont fonctionnellement et structurellement distincts des récepteurs aux protéines G. AdipoR1 est surtout présent dans le muscle et AdipoR2 dans le foie. L'adiponectine stimule la thermogenèse, l'oxydation des acides gras et exerce des effets hypolipémiants. Elle présente aussi des propriétés anti-diabétiques (en augmentant la sensibilité à l'insuline dans le muscle et le foie) et antiathérogènes (en partie liées à ses propriétés anti-inflammatoires sur l'endothélium) (Berg, Combs et al. 2002). Elle pourrait donc s'avérer très intéressante sur le plan thérapeutique chez le patient présentant un syndrome plurimétabolique.

	Adipocyte	**Tissu adipeux**	**SVF**	**Dépôt adipeux**	**Obésité**	**Perte de poids**
TNF-α	+		+++	Indifférent	↗	↘
PAI-1	+		++	Viscéral	↗	↘
IL-6	+		++	Viscéral	↗	↘
CRP		+		Viscéral	↗	↘
SAA3		+		Viscéral	↗	↘

ASP				Viscéral	↗	↘
Leptine	+++			Sous-cutanée	↗	↘
Adiponectine	+++			Sous-cutanée	↘	↗
Résistine	+		++	Viscéral	↗	↘
Visfatine		+		Viscéral	↗	?
VEGF	+		++	Viscéral	↗	↘
AGT		+		Viscéral	↗	↘
AT-II		+		Viscéral	↗	↘
MCP-1	+		++	Viscéral	↗	↘
MIF	++		++	Indifférent	↗	↘

Le site intratissulaire de production des différents facteurs sécrétés (adipocytes *versus* fraction stroma vasculaire) est indiqué quant il est connu, à défaut la production est attribuée au tissu adipeux. La contribution des différents dépôts adipeux est également indiquée. TNF-α : *Tumor Necrosis Factor alpha* ; PAI-1 : Plasminogène Activateur-Inhibiteur-1 ; IL-6 : Interleukine 6 ; CRP : *C Reactive Protein* ; SAA3 : *Serum Amiloid* ; ASP : *Acylation Stimulating Protein* ; VEGF : *Vacular Endothelial Growth Factor* ; AGT : Angiotensinogène ; AT-II : Angiotensine II ; MCP-1 : *Monocyte Chemoattractant Protein* 1;MIF : *Macrophage Inhibitory Factor* ; SVF : fraction stroma vasculaire.

Table 5 :Principaux facteurs sécrétés par le tissu adipeux et leur évolution en fonction de la masse grasse, d'après (Pégorier 2007)

Au cours de la dernière décennie, un nombre considérable d'études a démontré que le tissu adipeux sécrétait de très nombreuses protéines, peptides et facteurs lipidiques agissant de façon autocrine, paracrine ou endocrine, contrôlant non seulement son activité mais également le métabolisme énergétique dans sa globalité.

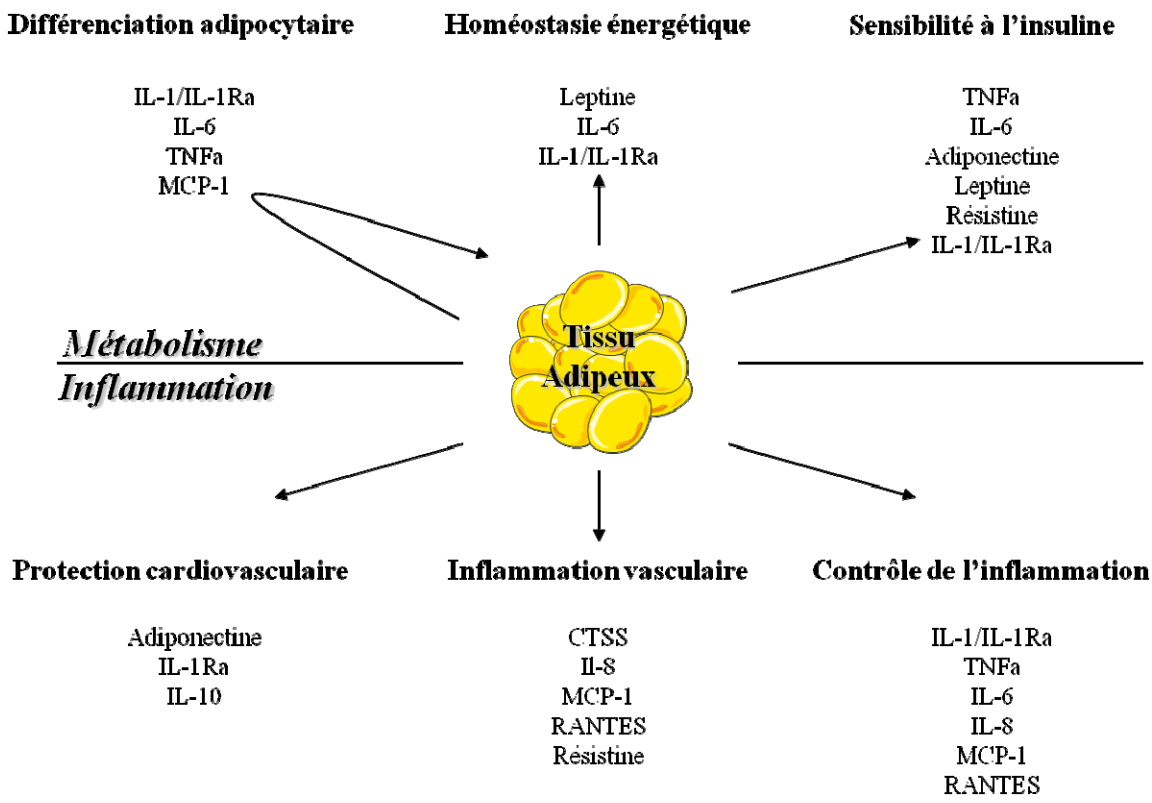


Figure 3 : Principales fonctionnalités des principaux biomolécules secrétés par le tissu adipeux(d'après (Juge-Aubry, Henrichot et al. 2005))

Les approches expérimentales ont fait évoluer considérablement la vision du tissu adipeux. Ces découvertes majeures ont permis de reconsidérer la fonction de ce tissu non plus comme un organe « passif », mais comme un tissu jouant un rôle central dans le développement de nombreuses pathologies. Le tissu adipeux viscéral sécrète beaucoup plus de facteurs pro-inflammatoires (TNF-α, IL-6, VEGF, AGT, AT-II, PAI-1, CRP, résistine, etc.) et de corticoïdes (dus à l'expression de la 11β-HSD1) que le tissu sous-cutané et ceci, aussi bien dans les modèles animaux que chez l'Homme. Or ces facteurs déversés directement dans la veine porte contribuent à amplifier l'état inflammatoire et à altérer le métabolisme hépatique (augmentation de la production de glucose, diminution de la sensibilité à l'insuline…). Il est donc sans doute important d'étudier l'expression génique de ce tissu afin de mieux comprendre son fonctionnement.

1.3 Sources de données et enjeux

Les progrès rapides de la biotechnologie et l'évolution de l'informatique sont à l'origine d'une évolution exponentielle des ressources biologiques disponibles et mises à la disposition des chercheurs (Birkland and Yona 2006; Mudunuri, Che et al. 2009). Face à la multiplication de ces ressources biologiques, l'extraction et l'exploitation de l'information devient de plus en plus complexe. Mais cette étape est primordiale afin de faire face aux défis actuels et être capable de comprendre les systèmes biologiques qui sont complexes et multifactoriels. Ce processus nécessite une combinaison d'informations biologiques diverses pour répondre à des problématiques nécessitant d'une part des données hétérogènes mais complémentaires et aussi des connaissances transversales afin d'être capables d'assembler ces données d'une manière cohérente et produire ainsi des modèles et des résultats pertinents (Shah, Huang et al. 2005; Lee, Pouliot et al. 2006).

Nous présentons dans le chapitre qui suit le matériel biologique utile pour aboutir à des données exploitables et expliquons l'acheminement et le processus qui permettent à partir des données biologiques l'obtention d'une quantification informatique de ces données biologiques. Cette représentation numérique des ressources informatique est le point de départ de toute analyse en bioinformatique.

Chapitre 2

Des données biologiques aux données transcriptomiques

Les biopuces représentent une des nouvelles technologies émergentes de la recherche biomédicale. L'application la plus répandue des biopuces concerne actuellement l'étude du transcriptome dans le cadre d'études différentielles, d'étude de profil génique ou d'études pronostiques ou prédictives. L'*apprentissage automatique (Trevor Hastie 2001; Bishop 2006)* joue un rôle majeur dans l'extraction des connaissances à partir de ces données. Ce domaine de recherche est à l'intersection de l'intelligence artificielle et des statistiques. Son essor n'a cessé de croître au cours des vingt dernières années. Il a entre autres pour objectif l'analyse des propriétés et la conception d'algorithmes qui permettent d'approximer des fonctions. Quand les co-domaines de ces fonctions sont finis, on parle de *classeurs*. Les applications de ces algorithmes en biomédecine sont innombrables (Golub 1999; Alizadeh, Eisen et al. 2000; Shipp, Ross et al. 2002). Quand les *exemples d'apprentissage* sont des points dans R^n, les approches classiques de régression aussi bien que celles issues de l'apprentissage statistique comme les réseaux de neurones ou les machines à vecteurs de supports (SVM) sont utilisées (Brown, Grundy et al. 2000; Dudoit 2002; Michael C. O'Neill and Song 2003). Lorsque les exemples sont structurés, on a recourt à des approches spécifiques d'apprentissage automatique d'*arbres ou règles de décision* (Breiman, Friedman et al. 1984). Un des verrous scientifiques actuels est de concevoir des algorithmes qui produisent des classeurs qui s'accommodent de faibles nombres d'exemples en regard du nombre de descripteurs. De ce fait, avec les données transcriptomiques, nous sommes confrontés au problème connu sous le nom du « fléau de la dimension »[2] puisque nous disposons de peu d'exemples et de milliers de descripteurs (les

[2] curse of dimentionality en anglais

valeurs d'expressions des gènes). Dans ce qui, suit nous introduisons les concepts biologiques nécessaires à la compréhension du principe des puces à ADN. Ensuite, nous décrivons les puces à ADN et leurs principes de préparation et enfin, nous présentons les méthodes d'apprentissages utilisées pour l'analyse des données issues de ces puces à ADN.

2.1 De l'ADN à l'homme

Le code génétique humain est construit à partir de molécules d'acide désoxyribonucléique (ADN). L'ADN est une molécule constituée de deux chaînes complémentaires enroulées en hélice. Chacune d'elles est formée par l'alternance d'un groupe phosphate et d'une molécule de sucre, le désoxyribose. À chaque molécule de sucre est fixée une base azotée parmi quatre possibles (Adénine, Thymine, Cytosine et Guanine). La complémentarité des deux brins d'ADN réside dans l'appariement spécifique des bases : l'adénine d'une chaîne ne peut se lier qu'avec la thymine de l'autre, la cytosine qu'avec la guanine. Le génome humain a entre 2,8 et 3,5 milliards de paires de bases. Les gènes sont des séquences de centaines ou de milliers de paires de bases qui fournissent les modèles pour toutes les protéines, dont le corps a besoin de produire. Le nombre total des gènes n'est pas connu, mais les estimations varient de 30000 à 120000. Ces gènes sont emballés en paquets appelés chromosomes. Tout être humain a 23 paires de chromosomes. Les 46 chromosomes sont localisés dans les noyaux des cellules. Presque chaque cellule du corps contient le code ADN complet pour produire un être humain. Chacune des cellules se différencie en obéissant à un certain nombre d'instructions dans l'ADN. Et de là résultent les cellules du sang, des muscles, des os, des organes internes. Le corps humain est construit à partir de 100 billions de ces cellules.

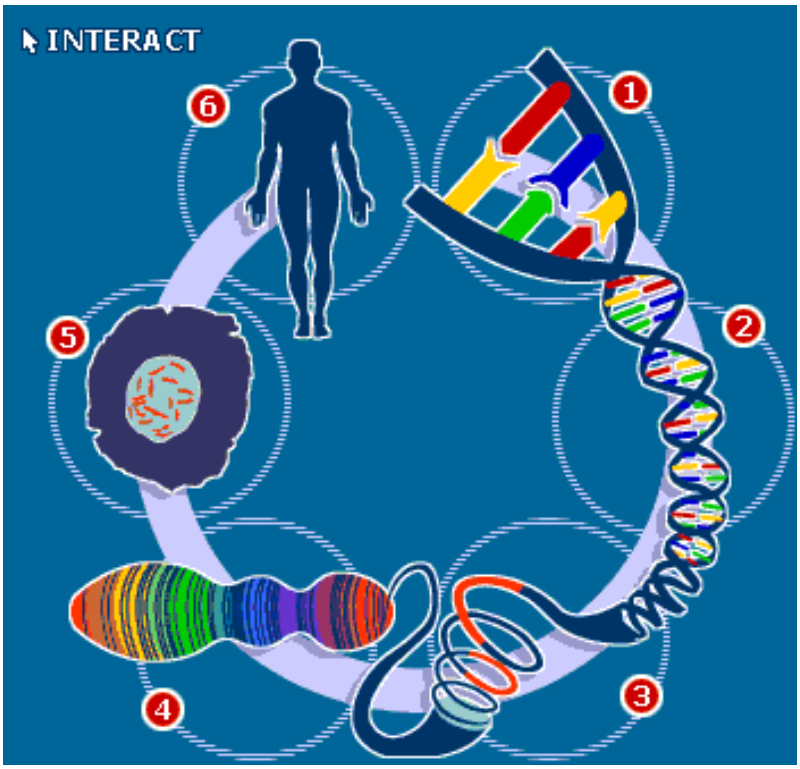


Figure 4: De l'ADN à l'homme. 1:paire de base ; 2 : ADN double Hélice ; 3 : gènes 4 ; chromosomes ; 5 : cellules ; 6 : corps humain.(source:bbc.co.uk, human genome project)

2.2 La biologie à haut-débit

Les avancées dans le domaine de la biologie et celui de l'informatique sont à l'origine de l'augmentation exponentielle des données biologiques disponibles de nos jours. Cela est à l'origine de l'expansion et du développement des sciences dites « omiques » : génomique, transcriptomique, protéomique et l'interactomique.

2.2.1 La génomique :

La génomique vise principalement à comprendre la complexité des interactions des gènes entre eux et avec leur environnement. Elle s'intéresse à l'étude du matériel génétique d'un individu ou d'une espèce qui est encodé dans son ADN. Ces molécules sont des polymères de petites molécules de base, les acides nucléiques, qui diffèrent entre elles par une partie appelée base

azotée. Il existe quatre types de bases azotées dans l'ADN, généralement notées A (adénine), T (thymine), C (cytosine) et G (guanine).

La base de données GOLD[Genomes On Line Database] (Liolios, Mavromatis et al. 2008) fournit des informations sur les projets de génomes et métagénomes dans le monde entier. En septembre 2007, GOLD contient des informations sur plus de 2900 projets de séquençage, dont 639 ont été achevés et leurs données sur les séquences déposées dans les bases de données publiques. Comme pour le génome, il existe aussi trois bases de données biologiques de séquencesnucléiques : GenBank (Benson, Karsch-Mizrachi et al. 2008) aux USA, EMBL (European Molecular Biology Laboratory) (Cochrane, Akhtar et al. 2007) et DDBJ (DNA Data Bank of Japan) (Sugawara, Ogasawara et al. 2008). Grâce à la collaboration internationale de ces trois bases de données depuis 1982, il existe une base internationale commune INSD
(International Nucleotide Sequence Databases) (Mizrachi 2008) où les séquences nucléiques sont échangées chaque jour pour assurer une synchronisation des informations. Les laboratoires ou les centres de séquençage envoient les séquences à l'aide des outils de soumission Bankit pour GenBank, Webin pour EMBL et Sakura pour DDBJ.

2.2.2 **La transcriptomique**

La transcriptomique s'intéresse aux gènes transcrits, c'est-à-dire aux ARN. L'ADN est répliqué de façon identique dans toutes les cellules de notre organisme. La présence de l'ARN, au contraire, évolue avec le temps, suivant la fonction de la cellule, etc. Il est traduit par la suite en protéines dans la cellule. La transcriptomique mesure l'expression des gènes, c'est-à-dire la quantité d'ARNm présente dans une cellule à un instant donné. Elle étudie aussi les mécanismes de transcription et d'épissage.

Il existe des bases de données spécialisées proposant un accès aux données publiques d'expression des gènes. Les plus connus sont: la Gene Expression Omnibus (GEO) (Barrett, Troup et al. 2007) gérée par le National Institute for Biotechnology Information (NCBI), ArrayExpress (Parkinson, Kapushesky et al. 2007) gérée par l'Institut Européen de Bio-informatique (EBI) et la Stanford Microarray Database (Ball, Awad et al. 2005).

2.2.3 **La protéomique**

Lors de la traduction, une machinerie cellulaire, le ribosome, décode l'ARNm par triplets de nucléotides, les codons, et fait correspondre à chaque codon un acide aminé. Il en résulte un polymère d'acides aminés appelé aussi polypeptide. Le polypeptide résultant de la traduction d'un ARNm se replie dans l'espace pour former une protéine. La protéomique permet de mettre en relation la séquence du génome et le comportement cellulaire. Son but est l'étude des produits protéiques dynamiques exprimés à partir du génome et leurs interactions à un moment donné ou sous certaines conditions environnementales. Leur connaissance et leur caractérisation sont donc essentielles à une meilleure compréhension de la cellule vivante.

Il existe des bases de données de données dédiées aux séquences protéiques. La base de données Uni-Prot (The UniProt Consortium 2008) est considérée en tant que base de données de référence. Cette base de données résulte de la fusion de SwissProt, TrEMBL et PIR.

2.2.4 **L'intéractomique**

L'intéractomique s'intéresse à étudier les interactions moléculaires dans la cellule. De ces interactions résulte l'ensemble des réactions formant le métabolisme d'une cellule. Cette notion est employée par défaut pour décrire l'ensemble des interactions protéine-protéine. Les protéines interagissent physiquement entre elles et avec d'autres molécules pour réaliser leur fonction ou activité biochimique. L'étude de l'activité biochimique des protéines ainsi que la connaissance des molécules avec lesquelles elles interagissent permettent de comprendre leurs fonctions cellulaires.

Il existe également des bases de données où sont stockées les interactions protéiques. Les bases de données les plus connues sont BIND (Biomolecular Interaction Network Database) (Bader, Betel et al. 2003),DIP (Database of Interacting Proteins)(Xenarios, Salwinski et al. 2002) et IntAct (Hermjakob, Montecchi-Palazzi et al. 2004). Il existe aussi d'autres bases de données d'interaction, par exemple, la base de données NPInter contient les interactions entre l'ARN non-codant et les protéines (Wu, Wang et al. 2006)

2.3 Les puces à ADN

Les puces à ADN, également appelées DNA arrays, DNA chips, DNA microarrays ou microarrays, ont été développées au début des années 1990. Schena et ses collègues à l'Université de Stanford ont publié le premier article sur les puces en 1995 (Schena, Shalon et al. 1995), cet article a été un catalyseur pour l'expansion de cette technologie dans le milieu académique et au sein des institutions privées. Depuis leur apparition, les puces à ADN sont devenues un outil majeur pour la recherche en biologie fondamentale et clinique. Parmi leurs nombreuses applications, citons la mesure de l'expression d'ARN messager (ARNm), la détection d'agents pathogènes, la détermination de génotypes, le reséquençage, l'étude de la méthylation de promoteurs et la mesure d'interactions entre biomolécules.

Les puces à ADN sont l'aboutissement des méthodes d'hybridation employées au cours des dernières décennies pour l'identification et le dosage des acides nucléiques présents dans les échantillons biologiques.

2.3.1 **Le principe des puces à ADN**

Le principe de fonctionnement des puces à ADN se fonde sur la propriété d'hybridation spécifique entre deux séquences complémentaires d'acides nucléiques. E.M. Southern a décrit en 1975 la première méthode de détection des acides nucléiques, en l'occurrence de l'ADN, par hybridation sur un support solide, méthode ainsi appelée Southern-blot (Southern 1975). Par la suite, cette méthode a été adaptée à l'étude des ARNm et a alors été appelée Northernblot. Les acides nucléiques analysés par Southern-blot ou Northern-blot sont d'abord immobilisés sur une membrane de nitrocellulose avant d'être hybridés avec une sonde nucléotidique marquée (radioactivité ou fluorescence), spécifique du gène cible. Le signal détecté sera alors proportionnel à la quantité de gènes cibles (Southern-blot) ou de son ou ses produits (Northern-blot) sur la membrane.

Les puces à ADN sont généralement réalisées sur des lames de verre d'un format compris entre un et quelques dizaines de cm^2. Ce support permet une approche miniaturisée et

de très haute densité (plusieurs dizaines de milliers de sondes par cm^2). Il est aussi adapté à l'analyse de faibles quantités d'ARN. Nous distinguons deux majeurs types de puce à ADN : Les puces à ADNc et les puces à oligonucléotides.

2.3.2 **Les puces à ADNc**

Le premier type de puce à ADN consiste en une lamelle de verre (identique à celle utilisée en microscopie traditionnelle) sur laquelle des milliers d'ADNc sont déposés à l'aide d'un micropipetteur robotisé. Grâce à cette technique, chacun des gènes est représenté par un seul point sur la lamelle. En général, deux échantillons d'ARN (sous forme d'ADNc obtenu par transcription inverse) sont co-hybridés sur la puce à ADNc. Les deux échantillons marqués par un fluorophore différent (Cy-3 vert ou Cy-5 rouge) s'hybrident simultanément avec les molécules complémentaires sur la puce. L'intensité du signal lumineux mesurée aux deux longueurs d'onde correspondant aux différents fluorophores est ainsi mesurée. Le rapport de fluorescence rouge/vert est alors déterminé et permet de comparer les taux d'expression relatifs de chacun des gènes pour les deux échantillons d'ADNc. Par convention, pour les stratégies à deux fluorophores, l'intensité d'hybridation est représentée par un dégradé de rouge ou de vert, si l'un ou l'autre des fluorophores prédomine, et de jaune quand les deux fluorophores sont de même intensité. L'intensité de la couleur est proportionnelle à la force du signal, le noir symbolisant l'absence de signal.(Figure 5).

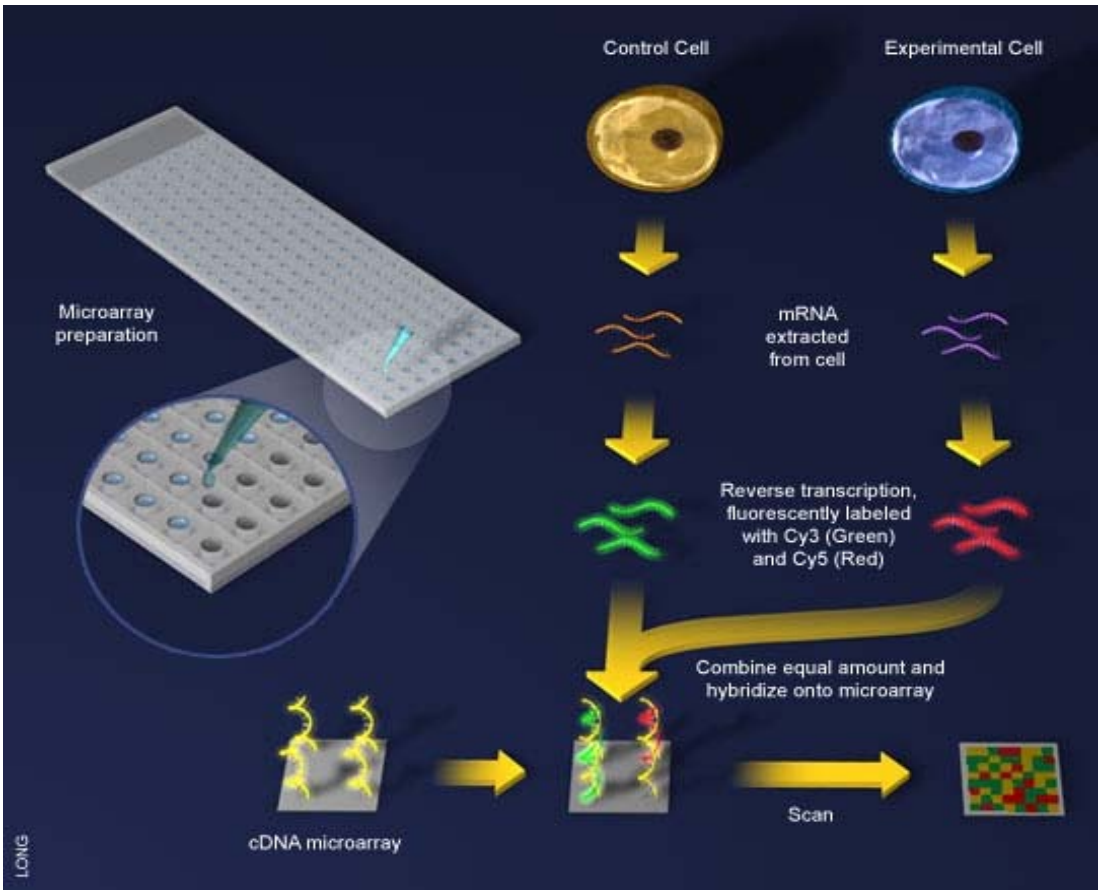


Figure 5 : Puces à ADNc préparées en parallèle à l'aide d'un micropipetteur robotisé qui dépose des ADNc sur la surface de la puce. Deux échantillons d'ARN provenant de différents tissus ou traitements sont marqués par des fluorophores différents (Cy-3 vert et Cy-5 rouge). La quantité relative de chaque gène est déterminée par le rapport d'émission de chaque fluorophore à des longueurs d'onde différentes (source : bioteach)

2.3.3 **Les puces à oligonucléotides**

Le second type de puce à ADN, est constitué d'oligonucléotides synthétisés directement sur un substrat solide par photolithographie. Dans ce procédé, une lumière dirigée sur des sites spécifiques de la puce, active la réaction d'oligo-synthèse (Chee, Yang et al. 1996; Lockhart, Dong et al. 1996). La synthèse d'un oligonucléotide de 25 paires de base occupe un carré de 20 µm x 20µm et contient plus de 107 copies de cette molécule.

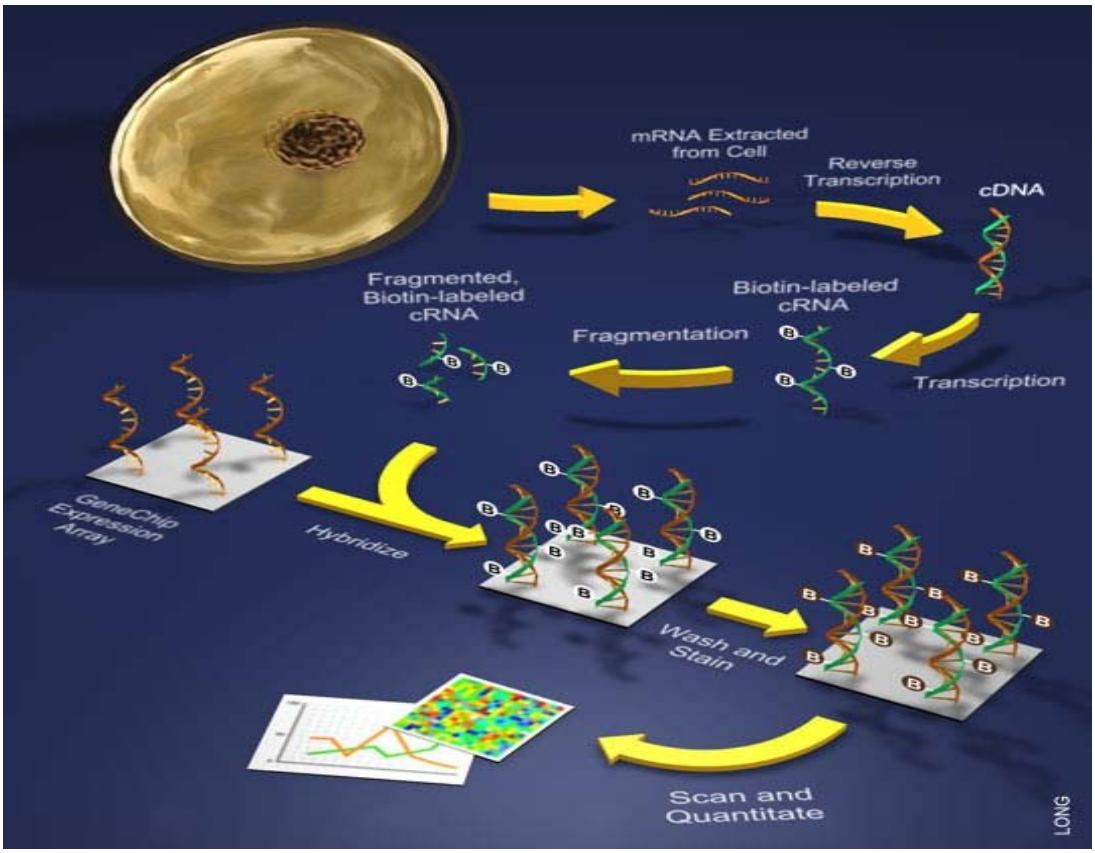


Figure 6: Utilisation des puces à Oligonucleotide. La puce à ADN est contenue dans une plaquette de plastique contenant une chambre d'hybridation. (Source : bioteach).

La surface d'une puce est d'environ 1.28cm², et peut contenir 400 000 oligonucléotides différents. Une puce à ADN destinée à des études d'expression contient pour chaque gène un ensemble d'oligonucléotides mimant la séquence du gène, souvent choisis dans sa région 3', réduisant ainsi les risques d'hybridations croisées avec des séquences homologues de ce gène. Des oligonucléotides, dont la séquence varie pour une seule base, sont également ajoutés, ce qui permet de confirmer que le signal obtenu pour chacun des gènes est bien spécifique. Contrairement à la puce à ADN décrite plus haut, celle produite par ce procédé permet l'hybridation d'un seul échantillon marqué à la fois).

2.3.4 **Transformation et gestion des données issues des puces à ADN**

Une fois les puces produites, l'étape suivante consiste à récupérer les signaux dans un format numérique analysable par ordinateur. Cette numérisation est réalisée avec des scanners de haute précision adaptés aux marqueurs utilisés. Les images obtenues sont ensuite traitées par des logiciels d'analyse d'images qui permettent de quantifier les signaux pour chaque fluorophore, mais aussi de relier chaque sonde à l'annotation correspondante (nom du gène, numéro de l'ADNc utilisé, séquence de l'oligonucléotide, etc.).

Si les biopuces fournissent simultanément un grand nombre d'informations, de nombreuses sources de variabilités ou d'erreurs sont introduites à chaque étape de l'expérience biologique. Par exemple, les propriétés des fluorochromes sont différentes, ou les conditions de température et d'humidité, qui influencent grandement l'étape d'hybridation, peuvent varier. On peut également rencontrer des problèmes de saturation des logiciels d'acquisition, etc. Toutes ces sources d'erreurs expérimentales introduisent une variabilité technique qui masque la variabilité biologique. Le rôle des étapes de prétraitement est d'extraire cette variabilité biologique (Ho, Stefani et al. 2008).

La grande diversité des technologies, des protocoles et des traitements des données rend d'ailleurs la comparaison des différentes études parfois périlleuse. Pour faciliter leur comparaison, la société savante Microarray Gene Expression Data Society (MGED) a créé un format standard de présentation comportant les informations nécessaires minimales pour décrire explicitement les données et en permettre la comparaison. Ce format MIAME (minimal information about microarray experiment) contient la description des séquences utilisées pour chaque sonde, la description des protocoles d'amplification et de marquage, l'annotation des échantillons utilisés, les données brutes, etc. Plusieurs standards ont aussi récemment été proposés par le Microarray Quality Control Consortium.

Les données biopuces prennent la forme d'une matrice d'expression. Chaque colonne correspond au profil d'expression d'un exemple sur l'ensemble des gènes mesurés. Chaque ligne correspond au profil d'expression d'un gène sur l'ensemble des exemples ou patients utilisés.

Chaque cellule de la matrice est une valeur continue qui représente le niveau d'expression d'un gène sur un exemple.

La gestion de cette grande quantité d'informations produites par ces technologies représente un défi énorme pour les biologistes. L'un des facteurs limitants est donc la capacité d'analyse de l'information. Pour cela, de nouvelles méthodes et outils informatiques sont nécessaires pour réaliser les études relatives aux expérimentations biologiques.Ainsi, il est primordial de rendre ces données exploitables afin d'être capable d'extraire des connaissances pertinentes à partir de cette masse de données. Un autre point important est celui de l'intégration de données, puisque ces sources biologiques sont de natures différentes et nécessitent des outils capables d'organiser les données, rendre les différentes sources de données exploitables et facilement interrogeables par les biologistes.

2.4 Exploitation des données biologiques

Lors de la réalisation d'une nouvelle expérimentation, la principale tâche pour rendre les données exploitables est celle de l'annotation d'un génome. Cette opération permet de caractériser la fonction des gènes, de leurs protéines et des processus biologiques associés. Pour cette tâche, trois ressources d'annotations génomiques sont incontournables pour réaliser ce travail : Gene ontology, KEGG ontology et NCBI.

2.4.1 **Gene Ontology**

Gene Ontology (The Gene Ontology Consortium 2008) est considéré comme la plus importante des ressources généralistes d'annotations génomiques. Cette ressource d'annotation est construite autour d'un noyau ontologique reliant plus de 25800 catégories terminologiques avec définitions (pour la version datant de Septembre 2008). Ces catégories sont réparties en trois thématiques : processus cellulaires, éléments structuraux et fonctions moléculaires (Ashburner, Ball et al. 2000). Une des limitations de GO est l'ignorance des homologies existantes entre des séquences géniques appartenant à des organismes différents.

Amigo (Carbon, Ireland et al. 2009) est le portail principal qui permet d'accéder à GO. Il contient notamment de nombreuses références croisées vers d'autres systèmes d'information.

2.4.2 **KEGG**

KEGG (Kanehisa, Araki et al. 2008) est une base de données des systèmes biologiques répertoriant des informations biomédicales diverses comme la structure détaillée des interactions moléculaires constituant les voies métaboliques et la description de la régulation dans les différents organismes.
KEGG regroupe quatre bases principales :

- PATHWAY : base de voies métaboliques, de régulation de gènes et d'expression du signal.
- GENES : base de données génomique.
- LIGNAD : base constituée de données sur les composés, les polysaccharides, les réactions et les molécules thérapeutiques.
- BRITE : regroupe tous les objets présents dans KEGG sous forme de hiérarchies. Ceci est d'une grande utilité pour accéder rapidement à un objet donné, mais aussi pour naviguer entre les différentes classes d'objets.

Un des atouts de cette ressource est le niveau de détail avec lequel sont modélisées les interactions moléculaires constituant les voies métaboliques et de régulation. Un langage de balisage spécialement conçu, appelé KEGG Markup Language (KGML), est utilisé pour représenter d'une manière opérationnelle les voies modélisées, permettant ainsi de réaliser diverses opérations avec les cartes KEGG.

2.4.3 **Données du National Center for Biotechnology Information (NCBI)**

NCBI comporte des ressources génomiques diverses (systèmes d'identification des séquences, annotations, données phylogénétiques, polymorphismes, références bibliographiques, etc.). Parmi ces ressources, une des plus importantes est Entrez Gene (Maglott, Ostell et al. 2007) qui permet l'identification des séquences géniques. Cet outil est

indispensable pour relier les différentes informations disponibles sur les gènes dans les bases biomédicales. Entrez Gene permet aussi de synthétiser l'information disponible dans les autres ressources de NCBI sur les séquences géniques répertoriées.

2.5 Données utilisées dans le cadre de nos analyses

Nous présentons dans ce qui suit les bases de données ayant servi pour notre étude. Ces ressources sont utilisées dans plusieurs études cliniques et biologiques. L'équipe 7 (nutrition et obésité) de l'UMRS872 du Professeur Clement s'intéresse à l'étude des mécanismes de l'induction pondérale précoce, à l'identification des processus d'adaptation tissulaire en réponse aux variations du statut nutritionnel, et à la caractérisation des biomarqueurs et des prédicteurs moléculaires des situations biocliniques liées à l'obésité. Cet axe de recherche se situe complètement à la croisée des chemins entre plusieurs disciplines, la clinique, la biologie cellulaire, la génomique fonctionnelle et la bioinformatique.

Les bases de données de l'obésité sont issues des différents projets de recherche dans lesquels est impliquée notre équipe. Les bases de données du cancer sont des bases publiques disponibles en ligne.

2.5.1 **Données obésité**

2.5.1.1 **Base VLCD**

Les sujets de l'étude sont 39 femmes, praguoises, adultes et non ménopausées, en bonne santé avec un I.M.C (Indice de masse corporelle)> 30 à l'origine, ne suivant aucune thérapie médicamenteuse régulière et pendant la période d'inclusion dans le protocole, qui est de type long, soit 30 jours.

Nous disposons pour chaque patient, à la fois de données cliniques et de puces à ADN.
La table obes39cli regroupant les données cliniques contient les mesures suivantes : la taille, le poids, L'Indice de masse corporelle, la glycémie, l'insulinémie, la masse grasse et l'indice d'insulino-sensibilité (quicki). Toutes les données sont de type numérique, les colonnes représentent les attributs cliniques et les lignes désignent les patients.

Un exemple d'entrée dans la table est donné par la Table 6.

Table 6 : Exemple de données cliniques (BD obésité)

	Taille (cm)	PoidsJ1 (Kg)	Glyc.J1 (mmol/l)	Insul.J1 (µU/ml)	MG.J1 (Kg)	IMC.J1 (kg/m²)	Quicki.J1
Patient 1	**166**	**102**	**4.22**	**1.75**	**37**	**36.9**	**0.46**
Patient 2	**156**	**109**	**4.52**	**1.85**	**36.30**	**44.8**	**0.41**
…	**…**	**…**	**…**	**…**	**…**	**…**	**…**
Patient n	**176**	**114**	**4.32**	**1.89**	**39.10**	**36.8**	**0.36**

La table obes39Exp contient l'expression de 39727 gènes, aucune sélection n'a été faite puisqu'on ne dispose pas de liste de gènes biologiquement pertinents pour l'instant. Les colonnes représentent les noms des gènes et les lignes désignent les patients. Les expressions des gènes sont des valeurs numériques, une valeur positive indique que le gène est surexprimé alors qu'une valeur négative indique que le gène est sous exprimé. La Table 7 donne un exemple d'entrée dans la table de données d'expression :

Table 7 : Exemple de données expressions (BD obésité)

	Exp.Gene1	Exp.Gene2	Exp.Gene3		Exp.Gene39727
Patient 1	**1.572108392**	**0.648162627**	**0.877616262**	**…**	**0.271666785**
Patient 2	**1.621083921**	**0.776162628**	**1.210839213**		**0.481626271**
…	**…**	**…**	**…**	**…**	**…**
Patient n	**0.271666785**	**1.029183921**	**0.616262812**		**1.021083921**

Un patient est considéré comme répondeur au régime si sa variation d'IMC est supérieure à 5,5% après un mois. Si ce n'est pas le cas, ce patient est considéré comme non répondeur. La base *VLCD* contient 28 patients répondeurs et 11 patients non répondeurs.

Table 8 : donnée de la base de l'obésité

	Base de données « VLCD »
Distribution des classes	**28 (répondeur) / 11 (non répondeur)**
Tables Cliniques	**7 attributs (taille, poids, IMC, glycémie, insulinémie, masse grasse, quicki)**

Tables expressions	39 727 / 80 gènes

2.5.1.2 **Base de la chirurgie de l'obésité (bypass)**

Cette base de données réunit 101 patients, massivement obèses (IMC 44 ± 10 kg/m^2, âgés de 40 ± 12 ans), recrutés et suivis dans le service de nutrition de l'Hôtel Dieu entre 2002 et 2006 dans le cadre d'une chirurgie de l'obésité. Les patients ont bénéficié d'une technique chirurgicale de type bypass, dont l'efficacité a été évaluée à 3 mois (perte3m) puis 6 mois (perte6m) (ces attributs prennent la valeur -1, si la perte de poids est inférieure au seuil fixé, soit un échec de la technique, et 1 si la perte de poids est supérieure au seuil fixé, soit un succès de la technique). La base de données comporte des variables cliniques et biologiques, recueillies avant l'intervention chirurgicale. Chaque patient est caractérisé par des variables biologiques réelles et des variables qualitatives (diabète oui/non, fumeur oui/non, etc.), caractérisant l'obésité et ses aspects cliniques et métaboliques ainsi que ses complications multiples.

Table 9: variables cliniques et biologiques de la base bypass

nom	Description	unité	Non Répondeur		Répondeur		
			moyenne	Ecart type	moyenne	Ecart type	P-Value
X.MFscWAT	% de macrophages dans le tissue adipeux sous-cutané		16.94	10.03	13.83	7.95	0.90
size.oWAT	Taille adipocytaire dans le tissue adipeux omentale		69.89	13.94	69.47	10.53	0.13
size.scWAT	adipocyte size in subcutaneaous adipose tissue		76.07	14.30	76.97	9.57	0.28
adiponectine	Adiponectine	µg/ml	6.30	2.90	6.40	2.68	0.14
agechir	Age		43.28	10.82	38.49	9.91	0.98
ALAT	Alanine aminotransferase	UI/L	38.11	29.21	38.11	24.74	0.00
Albu	Albumine	g/l	38.17	3.98	39.69	3.31	0.96
ASAT	aspartate aminotransferase		26.52	14.00	25.15	9.95	0.44
IMC	Indice de masse corporelle	kg/m²	49.50	7.90	47.60	6.42	0.83

calc	Calcium	mmol/l	2.30	0.09	2.29	0.08	0.49
CholT	Cholestérol	mmol/l	5.12	0.84	5.20	0.97	0.32
Foleryt	érythrocyte folates	ng/ml	283.77	120.31	294.02	118.43	0.34
MG	masse grasse	kg	59.17	13.07	56.60	11.13	0.66
ferr	Ferritine	ng/ml	112.11	144.00	124.36	183.57	0.30
Fgene	Fibrinogene	g/l	4.05	0.94	4.06	0.84	0.03
GGT	Gamma glutamyl transpeptidase	UI/l	46.35	37.98	56.27	77.36	0.60
G0	Glycémie	g/l	1.21	0.44	1.14	0.53	0.48
HDL	HDL cholestérol	mmol/l	1.31	0.44	1.26	0.40	0.46
Taille	Taille	cm	167.94	8.79	169.67	9.41	0.67
Hcys	Homocysteine	µmol/l	9.59	3.56	8.68	2.59	0.86
Ins0.IRMA	Insuline	UI/l	16.82	11.56	16.95	10.90	0.05
IL6	Interleukine 6	pg/ml	3.67	2.39	3.23	1.48	0.75
Fer	Fer	mmol/l	13.56	5.56	13.06	4.78	0.38
MM	Masse maigre	kg	62.93	9.57	63.21	12.00	0.09
Leptine	Leptine	ng/ml	58.24	24.69	55.68	21.97	0.43
oroso	Orosomucoide	g/l	0.96	0.25	0.98	0.17	0.31
Prealb	Prealbumine	mg/l	248.68	58.50	253.14	47.78	0.33
DER.m	Dépenseénergétique au repos	kcal/24h	2254.02	486.64	2281.87	460.37	0.24
Sele	Selenium	µmol/l	1.12	0.21	1.05	0.17	0.94
SAApop	serum amyloid A	µg/ml	26.45	14.86	17.74	11.18	1.00
Folser	serum folates	ng/ml	5.81	3.19	5.57	2.44	0.34
TSH	thyroid-stimulating hormone	UI/l	2.53	1.61	2.30	1.36	0.57
trigly	Triglycerides	mmol/l	1.77	0.89	1.68	0.88	0.41
A.umol	vitamine A	µmol/l	1.93	0.61	1.89	0.50	0.29
B1	vitamine B1	nmol/l	185.80	56.46	186.77	46.92	0.08
B12	vitamine B12	ng/l	390.51	156.36	370.30	124.58	0.53
D	vitamine D	ng/ml	13.45	5.68	15.40	7.99	0.85
E	vitamine E	µmol/l	28.29	8.31	27.68	7.65	0.31
Poids	Weight	kg	140.08	28.15	137.75	26.76	0.34
Zinc	Zinc	µg/dl	12.94	2.06	13.19	1.67	0.48

2.5.1.3 **Base Nugenob**

NUGENOB (NUtrients GENes for OBesity) est un projet européen qui a débuté en 2001 pour une durée de 4 ans. Ce projet a pour but de comparer les combinaisons phénotype-génotype de différents groupes de patients à celles observées dans la population générale. L'objectif du projet NUGENOB est d'élucider le rôle des interactions entre des variantes génétiques et l'alimentation, et en particulier la consommation de graisses. Cette étude permettra à terme de mettre en évidence des facteurs prédictifs pour les modifications de la composition corporelle observée dans la pathologie obèse.

Ces facteurs pourront être :

- des variantes génétiques ou des haplotypes identifiés dans des gènes dont l'expression est régulée par l'alimentation,
- l'expression différentielle de ces gènes dans le tissu adipeux,
- des facteurs de style de vie associés à l'obésité,
- des variations de certaines fonctions physiologiques observées durant le test,
- des interactions génotype-phénotype ou gènes-environnement, ou enfin la combinaison de ces différents facteurs.

Ce projet regroupe douze partenaires européens ayant chacun une tâche spécifique allant de la collecte des patients et des sujets témoins à l'identification des gènes et des variants fonctionnels. Notre contribution à ce projet concerne la mise en place de modèle prédictif de la perte de poids après un régime de dix semaines.

Les données de cette base concernent 54 sujets de sexe féminin recrutés dans le cadre du projet NUGENOB. L'objectif de ce programme européen est d'élucider le rôle des interactions entre des variantes génétiques et l'alimentation, et en particulier la consommation de graisses. L'objectif de notre étude est de déterminer la faisabilité de la mise en place d'un

modèle prédictif de la perte de poids suite à un régime pauvre en graisse de 10 semaines à partir de l'expression des gènes du tissu adipeux sous-cutané.

319 sujets ont été évalués pour la perte de poids après 10 semaines et ensuite divisés en deux groupes : «répondeur» (perte de poids entre 8 et 12 kg) et "non-répondeurs" (perte de poids inférieure à 4 kg) (Figure 1). 27 sujets de sexe féminin ont été choisis au hasard dans chaque groupe après une correspondance en fonction des paramètres cliniques comme l'âge, le poids, l'indice de masse corporelle (IMC), etc.

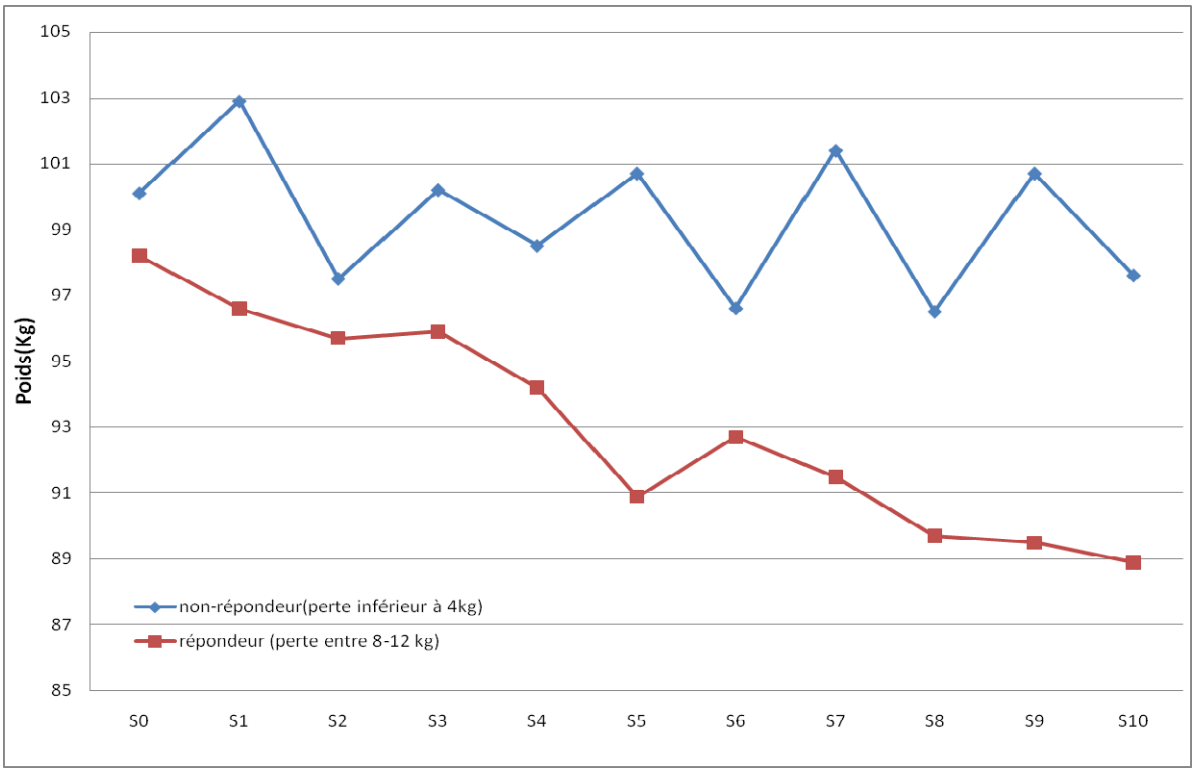


Figure 7 : évolution de la perte de poids des sujets de l'étude NUGENOB

Pour la sélection de patients qui a été retenue pour notre analyse, 54 puces à ADN Agilent ont été réalisées, ce qui correspond à une puce par sujet. Cependant, une puce d'une patiente du groupe « non répondeur » a été éliminé et ce, dû à la mauvaise qualité de celle-ci. La normalisation des données a été réalisée par la méthode « loess » (Smyth and Speed 2003).

14135 gènes étaient présents en commun sur toutes les puces. En plus des données biopuces, la base contient 1024 données cliniques.

Table 10 : données de la table Nugenob

	Base de données « Nugenob »
Distribution des classes	27 (répondeur) / 26 (non répondeur)
Tables Cliniques	1024 attributs cliniques
Tables expressions	14135 gènes

2.5.1.4 **Base Diogenes**

Dans le cadre de ce projet européen, 596 sujets ont suivi un régime hypocalorique de 8 semaines et ayant perdu 8% de leurs poids à la fin de ce régime ont été sélectionnés pour cette analyse. Parmi ces sujets, seulement 513 patients avaient les données nécessaires disponibles avant (temps : CID1) et après (temps : CID2) le régime. Un spécimen de graisse sous-cutanée abdominale (~ 1 g) a été obtenu par aspiration sous anesthésie locale. Les biopsies ont été stockées à -80 ° C jusqu'au moment de l'analyse. Pour cette analyse, nous avions besoin d'avoir des données de patients après 48h du début du régime (temps : CID1b) et cela afin d'avoir des modèles de prédiction basés sur une courte variation temporelle. Seulement 79 patients répondent à ce critère et parmi eux 67 ont une quantité d'ARN extraite suffisante pour l'extraction d'ARN. Ces patients proviennent de 3 centres de recrutement (Angleterre, Danemark et Hollande). La présente étude ciblant les femmes, nous avons donc retenu 44 sujets parmi les 79 qui étaient des femmes. Pour construire les groupes, nous avons sélectionné, comme pour Nugenob, la perte de poids (Kg) comme critère, mais nous avons remarqué une différence du BMI au temps CID1, ce qui nous a conduits, à choisir la variation en % de la perte de poids comme critère de sélection des groupes. Deux groupes ont été formés pour la variation de la perte de poids après 8 semaines : les « bons répondeurs » (sujets ayant perdu entre 13% et 17% de leur poids initial) et les « mauvais répondeurs » (sujets ayant perdu entre 8% et 10% de leur poids initial). 27 sujets répondent à ces critères au final, parmi eux 10 « bons répondeurs » et 17 « mauvais répondeurs ». Pour les sujets sélectionnés, les ARN totaux

obtenus à partir de biopsies prises à CID1 et CID1b ont été utilisés dans la présente étude prédictive. La Figure 21, décrit l'évolution de la perte de poids pour les deux groupes de l'étude tout au long du régime alimentaire.

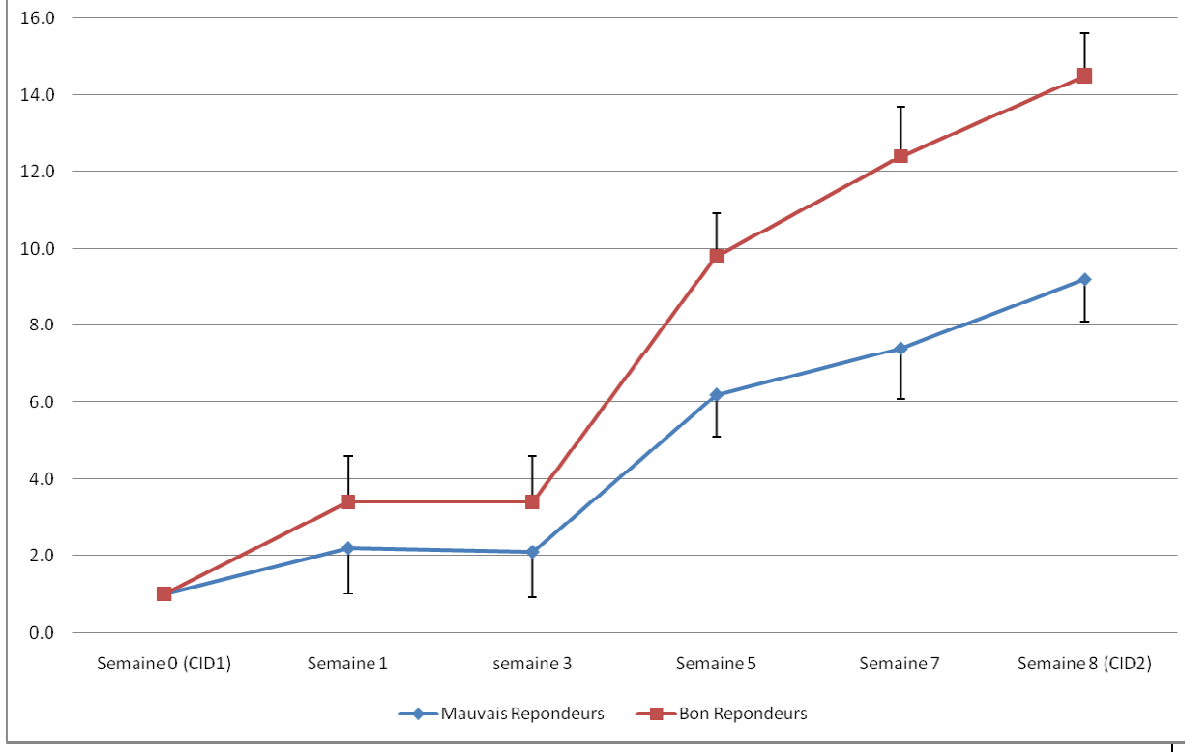


Figure 8: évolution de la perte de poids des sujets de l'étude DiOGenes

2.5.2 **Données cancer**

Nous avons utilisé dans notre analyse les données publiques du cancer du poumon de l'université de Harvard et de Michigan (Bhattacharjee, Richards et al. 2001) ainsi que les données de tumeur de cerveau de l'université de Massachusetts (Pomeroy, Tamayo et al. 2002).

2.5.2.1 **Données de la tumeur du cerveau**

La sélection que nous avons retenue dans la base de données de patients atteints de la tumeur du cerveau de l'université du Massachusetts regroupe 40 patients. Nous avons gardé seulement les patients ayant à la fois des données cliniques et des données d'expressions

disponibles simultanément. Dans cette sélection, 22 patients ont survécu après 5 ans et 18 sont décédés avant 5 ans. L'ensemble de données clinique regroupe le stade de la tumeur, l'âge au diagnostic, le sexe, la chimiothérapie (V=vincristine, C=cisplatin, Cx=cytoxan, VP=etoposide, CC=ccnu, Ca=carboplatin, P=procarbazine, M=methotrexate, T=thiotepa) et le sous-type de la tumeur (classique, Desmoplastique). Nous avons choisi la même codification présentée précédemment et nous avons opté pour coder les chimiothérapies et le sous-type par des variables booléennes. Dans cette base, les données biopuces regroupent l'expression de 7129 gènes.

	Base MIT
Distribution des classes	22 (survie) / 18 (non-survie)
Tables Cliniques	15 attributs (StageT, StageM, DiagAge, Sex,V, C, Cx, VP, CC, Ca, P, M, T, Subt.C, Subt.D)
Tables expressions	7 129 gènes

2.5.2.2 **Données cancer du poumon**

Pour la base de données du cancer de poumon de Harvard, nous avons sélectionné 43 patients ayant à la fois les données cliniques et les données d'expression disponibles et nous avons exclu tous les patients qui avaient des données manquantes ou ceux qui ont été censurés et dont le temps de survie était plus court que 5 ans. Dans notre sélection, 22 patients ont survécu après 5 ans et 21 sont décédés avant 5 ans. Pour la base de données Michigan, 93 patients sont retenus pour l'étude basée sur les mêmes critères que la base précédente. Pour chacune de ces deux bases, nous disposons de données cliniques et transcriptomiques. Les données cliniques regroupement l'âge, le sexe, la classification TNM et le stade du cancer. Pour la classification TNM, la lettre T (de l'anglais «tumor», tumeur) s'applique à la taille et à l'emplacement de la tumeur primitive, la lettre N («node», ganglion) indique si des cellules cancéreuses ont envahi les ganglions lymphatiques qui drainent des liquides dans la partie du corps où est située la tumeur et la lettre M («métastase», métastase) indique si le cancer s'est propagé à d'autres régions de l'organisme. Nous avons codé la classification TNM en utilisant trois attributs, par exemple T2M1N0 est codé (2, 1, 0) et lorsque nous avons M1, le code

correspondant sera (-1, -1, 1). Nous ont par ailleurs converti les stades du cancer en nombre (IA=1, IB=2, IIA=3, IIB=4, IIIA=5, IIIB=6, IV=7). La table de données d'expression contient le profil transcriptionnel de 3588 gènes.

Table 11 : récapitulatif des bases de données cancer

	Base de données « Cancer du poumon »	
	Harv43	Mich93
Distribution des classes	22 (survie) / 21 (non-survie)	26 (survie) / 67 (non-survie)
Tables cliniques	**6 attributs (Age, Sexe, T, N, M, Stade)**	
Tables expressions	3 588 gènes	

Maintenant que nous avons présenté les données biologiques et le processus pour rendre ces données exploitables informatiquement, nous allons présenter dans le chapitre suivant les aspects méthodologiques de la fouille de données avec une illustration des applications les plus courantes dans le domaine.

Chapitre 3

Aspects méthodologiques de la fouille de données biomédicales

Dans ce qui suit nous allons introduire le principe de fonctionnement de l'apprentissage automatique ensuite présenter deux grandes familles de méthodes avec un fort impact sur l'analyse et l'exploration des données biomédicales, il s'agit des méthodes d'apprentissage non supervisé et des méthodes d'apprentissage supervisées. Les méthodes d'apprentissage non supervisé sont utilisées pour la découverte de groupes dans les données, alors que les méthodes d'apprentissage supervisé disposent d'une subdivision des données en deux ou plusieurs groupes et leur tache consistent à trouver une bonne séparation entre les objets de chaque groupe. Nous finirons ce chapitre par une étude de cas qui reprend une analyse prédictive réalisée dans le cadre du projet NUGENOB.

3.1 Concept de l'apprentissage automatique

L'apprentissage automatique(Bishop 2006) est une discipline qui s'intéresse à l'extraction et l'exploitation automatique d'informations présentes dans les données. Le problème de l'apprentissage en général est de construire une procédure permettant d'associer une classe à un exemple. Ce problème se décline en deux variantes : l'approche supervisée et l'approche non supervisée. Dans la première, on connaît les classes possibles et on dispose d'un ensemble d'exemples déjà classés, servant d'ensemble d'apprentissage qui permettent de construire une règle de séparation entre les classes. Le problème est d'être capable par la suite d'utiliser la règle de séparation afin d'associer à tout nouvel exemple la classe la plus adaptée. Dans la seconde approche, les classes possibles ne sont pas connues à l'avance, et les exemples

disponibles sont non étiquetés. Le but est de regrouper dans un même groupe les exemples considérés comme similaires, pour constituer des classes.

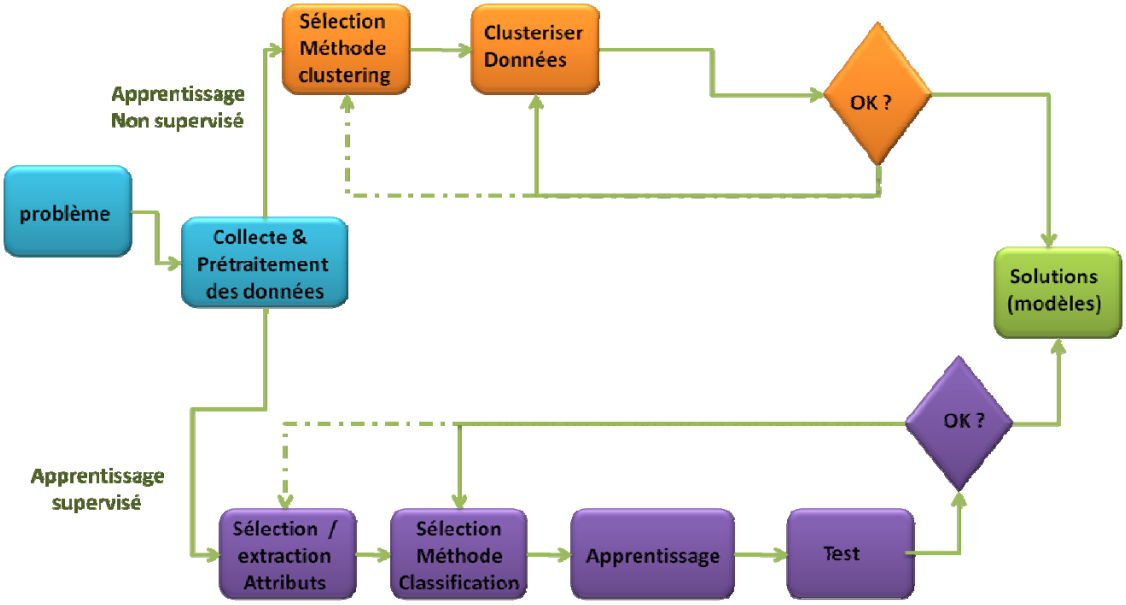


Figure 9 : cycle de l'apprentissage automatique, d'après (Kuncheva 2004)

3.2 Application des approches d'apprentissage non supervisé aux puces à ADN

Les approches non supervisées, ou analyses de classes (cluster analysis) sont des méthodes qui visent à regrouper les gènes et/ou les échantillons en fonction de leur ressemblance.

Dans ce type d'analyse, les groupes préalablement connus (type histologique, données cliniques…) ne sont pas pris en compte dans l'algorithme servant à classer les échantillons. Les deux mesures les plus fréquemment utilisées pour déterminer la similarité/proximité entre deux gènes/échantillons sont la distance euclidienne et le coefficient de corrélation de Pearson(Mount 2004). Différents algorithmes sont ensuite utilisés pour déterminer les groupes comme la classification hiérarchique, la classification par les nuées dynamiques (K-means), les cartes de Kohonen (self organizing map ou SOM). Ces méthodes sont utilisées quand aucune

information préliminaire n'est disponible sur les groupes d'échantillons ou pour analyser les profils moléculaires indépendamment de ces informations. Suivant l'algorithme utilisé, on peut ou non choisir le nombre de classes. Ce type d'approche est adapté à l'identification de nouvelles classes. Toutefois, déterminer l'implication ou la pertinence d'un tel groupe dans un contexte clinique ou biologique nécessite une grande expertise et demande une validation indépendante.

3.2.1 **Classification hiérarchique**

Le principe qui régit la classification hiérarchique consiste à regrouper les objets, dans le cadre de la transcriptomique. Il s'agit de gènes ou d'individus, sous la forme d'un arbre et cela soit d'une manière agglomérative (Lukasova 1979; Day and Edelsbrunner 1984; Yager 2000)ou bien d'une manière divisive (Marengo and Todeschini 1993). L'approche agglomérative part d'une configuration dans laquelle chaque objet constitue à lui seul un groupe, ensuite les paires de groupes les plus proches sont fusionnées jusqu'à ce qu'une condition d'arrêt est satisfaite. L'approche divisive quant à elle, est initialisée avec une configuration dans laquelle tous les objets sont placés dans un seul groupe et ensuite ce groupe est subdivisé itérativement jusqu'à ce qu'une condition d'arrêt est satisfaite (Figure 10). La construction des groupes se base sur deux mesures de similarité et qui sont dans ce cas des distances : une distance entre les entités (la distance euclidienne ou la corrélation) et une distance entre les groupes constitués (jonction simple, jonction moyenne ou jonction complète).

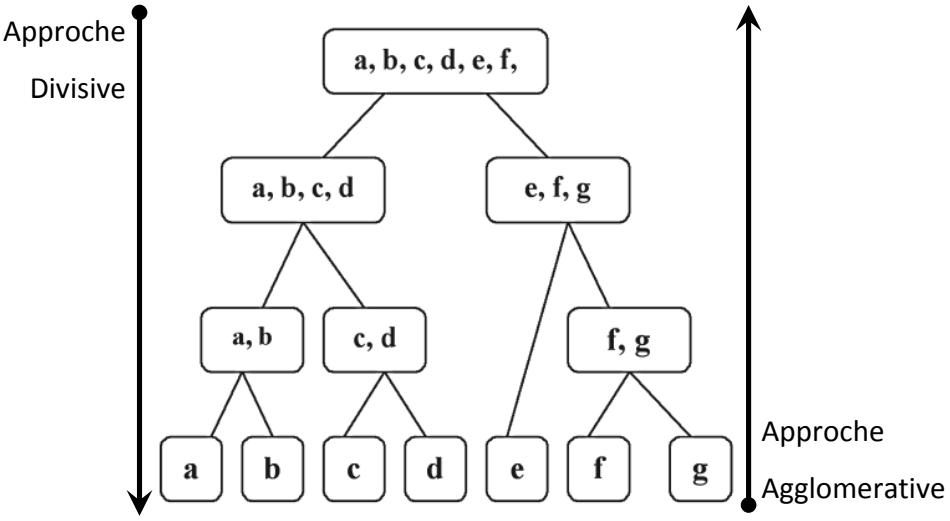


Figure 10: Les deux approches de classification hiérarchique.

Eisen et al. se servent de cette méthode pour proposer un mode d'interprétation visuelle des données de biopuces (Eisen, Spellman et al. 1998). Ils proposent une méthode de classification hiérarchique couplée à une colorisation du tableau de données qui met en évidence les proximités entre gènes et entre individus respectivement, en permettant une double classification des gènes et des individus. La métrique qui définit ces proximités est basée sur la corrélation. La Figure 11 , issue des résultats d'une étude conduite sur le lymphome (Alizadeh, Eisen et al. 2000), illustre cette méthode. Sur la gauche de la figure est représenté le dendrogramme pour les gènes, et sur le haut le dendrogramme des patients. Cette visualisation permet de regrouper les patients partageant des sous-types de pathologies communes. Parallèlement, les réseaux de gènes dont les niveaux d'expression sont propres à chacun de ces sous-types sont identifiés. Les zones rouges et vertes correspondent respectivement à une sur-expresssion et sous-expresssion des gènes considérés.

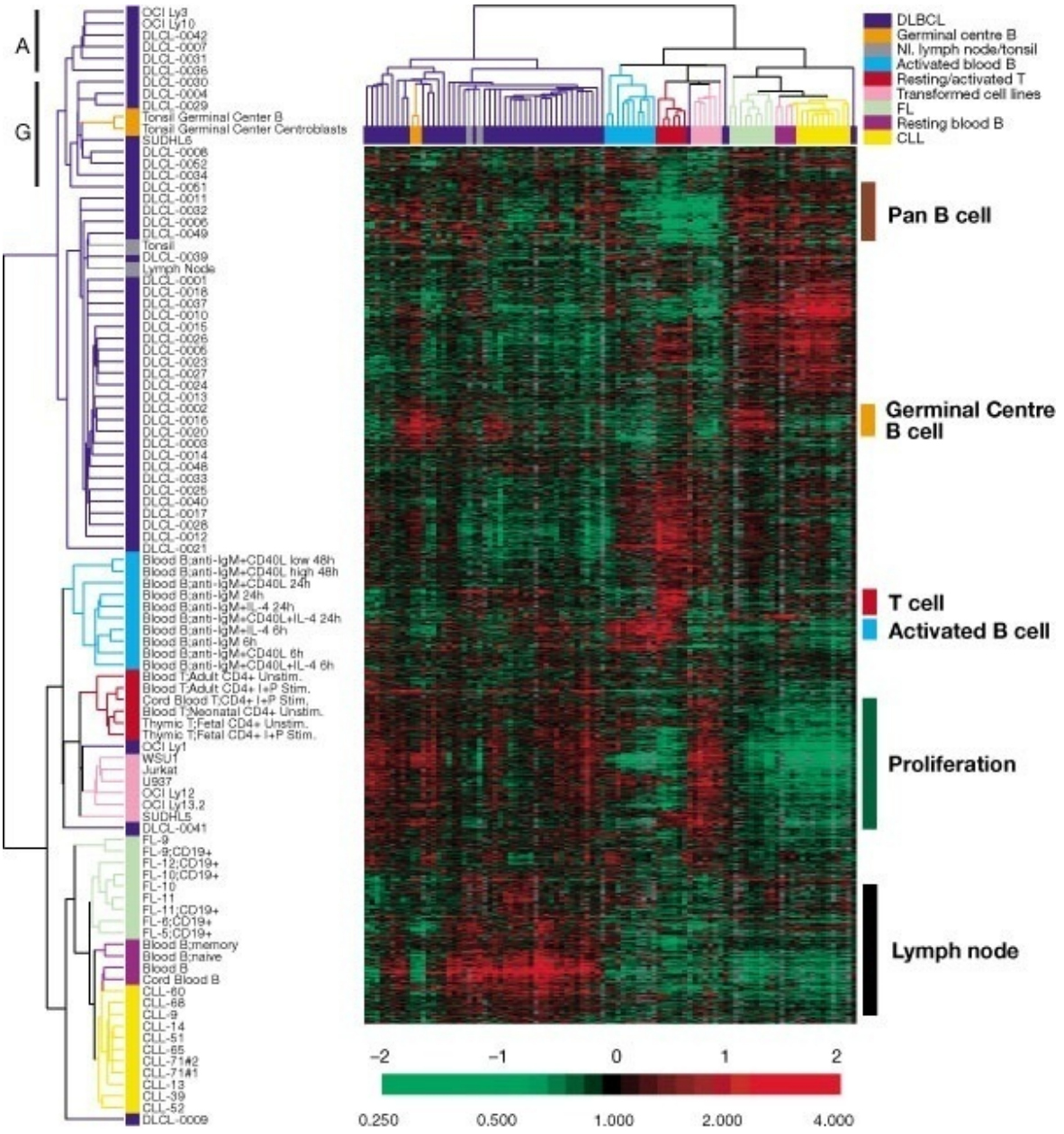


Figure 11: Exemple de visualisation de clusters hiérarchiques issu d'une étude sur le lymphome D'autres études similaires dans le domaine de la recherche sur le cancer on également utilisé ces méthodes (Mougeot et al., 2006, Nielsen et al., 2002, Ramaswamy et al., 2003, Welcsh et al., 2002, NHC Au, 2004, Makretsov et al., 2004)

3.2.2 **La classification par les nuées dynamiques (K moyennes)**

Cette méthode est très utilisée et son principe est relativement simple. Elle regroupe les objets selon un procédé non hiérarchique (Hartigan 1975). La première étape dans le processus de regroupement consiste à sélectionner K objets au hasard, chacun représentant la moyenne ou centroïde. Ensuite, les objets sont affectés aux groupes en regardant leur similarité aux centroïde préalablement définis selon une métrique choisie. Une fois l'ensemble des objets ainsi répartis dans les k groupes, les positions des k centroïdes sont recalculés. On réitère le processus jusqu'à ce que les positions des centroïdes restent fixes. Il est à noter que les résultats de cet algorithme sont sensibles à la position des k centroïdes initiaux.

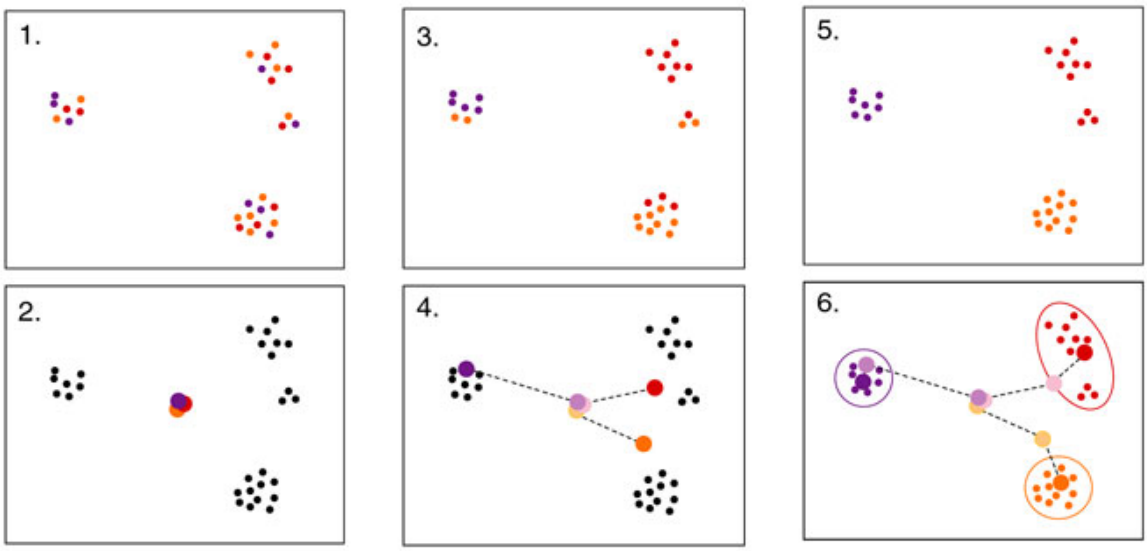


Figure 12 : K-moyenne pour le partitionnement des données d'expression génique (Gasch and Eisen, 2002. Genome Biol 3, 1–22). 1 : les gènes sont placés aléatoirement dans 3 groupes indiqués par les 3 couleurs. 2 : la moyenne du profil d'expression de chaque groupe de gènes est calculée comme centroïde (grands cercles), et les gènes sont réaffectés au centre auquel ils sont les plus proches. (4-6) les étapes 2 et 3 sont répétées jusqu'à ce que les centroïdes soient stables. Dans cette configuration les gènes sont affectés au groupe qui leur est proche.

Ces méthodes ont été utilisées en génomique dans différentes études (Tavazoie et al., 1999, Shai et al., 2003, Herwig et al., 1999)

3.2.3 **Les cartes auto-organisatrices**

Les cartes auto organisatrices (self-organizing maps) sont très utilisées pour l'analyse de données. Elles permettent de cartographier en deux dimensions et de distinguer des groupes dans des ensembles de données. On les désigne souvent par le terme carte de Kohonen du nom du statisticien ayant développé le concept en 1984 (Kohonen 1995).

Les SOM (Self Organizing Maps) sont un modèle de réseau de neurones artificiels qui va permettre de projeter les données non linéairement sur une grille : la carte de Kohonen. L'objectif de l'algorithme SOM consiste à placer sur la carte les objets de l'espace d'entrée (les gènes / les patients), tout en préservant les voisinages.

Cette méthode de quantification vectorielle permet de simplifier et de réduire la haute dimensionnalité des données d'expression(Wang, Delabie et al. 2002). Elle permet aussi de faciliter la visualisation de données complexes, ainsi que l'analyse de données de grande échelle (Tamayo et al., 1999, Toronen et al., 1999, Thalamuthu et al., 2006).

3.2.3.1 **Notations-Définitions**

Soit D l'espace des observations; les observations sont supposées quantitatives ou qualitatives et de grande dimension; on suppose que chaque observation est de dimension d. On suppose, par la suite que l'on dispose d'observations correspondant à N individus représentés par l'ensemble des couples $A = \{(\mathbf{z}_i, y_i); i = 1..N\}$ où l'observation est $\mathbf{z}_i$ et y_i l'étiquette de sa classe. Cette étiquette sera utilisée dans l'apprentissage supervisé (SVM).
La méthode de partitionnement cherche à déterminer une partition de D en N_{cell} sous-ensembles qui sera notée $P = \{P_1, ..., P_{N_{cell}}\}$. A chaque sous-ensemble P_c, on associe un vecteur référent $\mathbf{w}_c \in D$ qui sera le représentant ou le "résumé" de l'ensemble des observations de P_c. Par la suite, nous notons $W = \{\mathbf{w}_c; c = 1..N_{cell}\}$ l'ensemble des vecteurs référents. La partition P

de D peut être définit d'une manière équivalente avec la fonction d'affectation φ qui est une application de D dans l'ensemble fini des indices $I = \{1,2,...,N_{cell}\}$.

Dans le cas où il y a eu regroupement des sous-ensembles, nous avons défini une application surjective χ de I dans l'ensemble des indices $J = \{1,2,...,S\}$ où $1 \le S \le N_{cell}$. Si on utilise ces définitions, le sous-ensemble P_c est alors représenté par $P_c = \{\mathbf{z} \in D / \varphi(\mathbf{z}) = c, \chi(c) \in J\}$, (si $\chi(c) = 1$ alors $P = P_c = A$). On notera pas la suite l'ensemble des indices I_p des sous-ensemble pur sont tel que $I_p = \{c / \forall \mathbf{z} \in P_c, \chi(\varphi(\mathbf{z})) = c, vote(P_c) = y_c\}$. y_c est l'étiquette du vote majoritaire à 100% du sous-ensemble P_c en utilisant la fonction $vote$

3.2.3.2 **Cartes Topologique Mixtes**

Dans le cas où les données possèdent non seulement des attributs numériques mais aussi des attributs catégoriques, les cartes topologiques classiques ne sont plus applicables, nous utilisions dans ce cas la les carte topologiques mixtes, qui permet de travailler avec ce type de données. Dans la section suivante, nous présentons un modèle original de cartes topologiques dédié aux données mixtes avec la prise en compte des deux espaces réel et binaire en définissant un hyper-paramètre pour contrôler les variables quantitatives et qualitatives codées en binaire.

On suppose que l'on dispose dans le cas des cartes topologiques de la base d'apprentissage A sans les étiquettes $A = \{\mathbf{z}_i; i = 1..N\}$. Les observations $\mathbf{z}_i$ sont composées de deux parties: la partie numérique $\mathbf{z}_i^r = (z_i^{1r}, z_i^{2r}, ..., z_i^{nr})$ ($\mathbf{z}_i^r \in R^n$), et la partie binaire $\mathbf{z}_i^b = (z_i^{1b}, z_i^{2b}, ..., z_i^{mb})$ ($\mathbf{z}_i^b \in \beta^m = \{0,1\}^m$). Avec ces notations, une observation $\mathbf{z_i} = (\mathbf{z_i^r}, \mathbf{z_i^b})$ est de dimension $d = n + m$ (numérique et binaire).

Comme tout modèle de cartes topologiques, nous supposons que l'on dispose d'une carte discrète c ayant N_{cell} cellules structurées par un graphe non orienté. Cette structure de graphe permet de définir une distance, $\delta(r,c)$ entre deux cellules r et c de C, comme étant la longueur de la plus courte chaîne permettant de relier les cellules r et c, (voir figure 1). Le

système de voisinage est défini grâce à la fonction noyau K ($K \geq 0$ et $\lim_{|x|\to\infty} K(x) = 0$). L'influence mutuelle entre deux cellules c et r est définie par la fonction $K(\delta(c,r))$.

A chaque cellule c de la carte, est associé un vecteur référent $\mathbf{w_c} = (\mathbf{w_c^r}, \mathbf{w_c^b})$ de dimension d où $\mathbf{w}_c^r \in R^n$ et $\mathbf{w}_c^b \in \beta^m$. Par la suite, nous notons W l'ensemble des vecteurs référents constitués par les parties numériques et par la partie binaire.

L'algorithme d'apprentissage associé est dérivé de l'algorithme Batch de Kohonen dédié aux données numériques(Kohonen 1998) et de l'algorithme BinBatch dédié aux données binaires (Lebbah, Badran et al. 2000). Dans cet algorithme, l'indice de similarité et l'estimation des vecteurs référents sont spécifiques pour chaque partie de la base: c'est la distance euclidienne avec le vecteur moyen pour la partie numérique et la distance de Hamming et le centre médian pour la partie binaire.

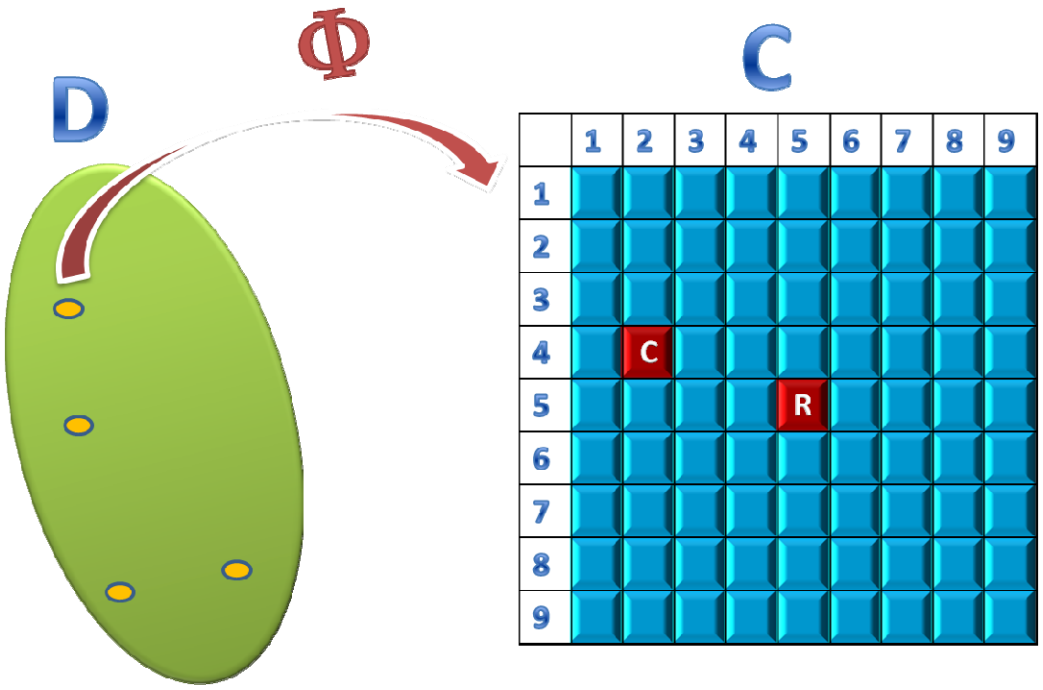


Figure 13: Carte topologique de dimension 9×9, ($\delta(c,r) = 4$).

φ est la fonction d'affectation de l'espace des données D dans l'espace de la carte C.

3.2.3.3 **Minimisation de la fonction de coût**

Comme dans le cas des cartes topologiques (Kohonen 1998) nous proposons de minimiser la fonction de coût suivante:

$$E(\varphi,W)=\sum_{\mathbf{z}_i\in App}\sum_{r\in C}K(\delta(\varphi(\mathbf{z}_i),r))\,\|\,\mathbf{z}_i-\mathbf{w}_r\,\|^2 \qquad (1)$$

Où φ affecte chaque observation $\mathbf{z}$ à une cellule unique de la carte C .

Dans cette expression, $\|\,\mathbf{z}-\mathbf{w}_r\,\|^2$ représente le carré de la distance euclidienne. Etant donné que, pour les données binaires, la distance euclidienne n'est rien d'autre que la distance de Hamming H , la distance euclidienne peut être réécrite: $\|\,\mathbf{z}-\mathbf{w}_r\,\|^2=\|\,\mathbf{z}^r-\mathbf{w}_r^r\,\|^2+H(\mathbf{z}^b,\mathbf{w}_r^b)$. Pour contrôler les deux parties des données (réelles et binaires), nous avons utilisé un hyper-paramètre F , qui respecte la propriété $0\le F\le 1$ pour pondérer les variables réelles et qualitatives codées en binaire. Cette pondération permet de pallier le problème d'échelle entre les variables binaires et réelles normalisées entre 0 et 1 . Ainsi, la distance est égale à :

$$\|\,\mathbf{z}-\mathbf{w}_r\,\|^2=(1-F).\|\,\mathbf{z}^r-\mathbf{w}_r^r\,\|^2+F.H(\mathbf{z}^b,\mathbf{w}_r^b).$$

Utilisant cette expression, la fonction de coût devient :

$$E(\varphi,W)=(1-F)\sum_{\mathbf{z}_i\in App}\sum_{r\in C}K(\delta(\varphi(\mathbf{z}_i),r))D_{euc}(\mathbf{z}_i^r,\mathbf{w}_r^r)$$

$$+F\sum_{\mathbf{z}_i\in App}\sum_{r\in C}K(\delta(\varphi(\mathbf{z}_i),r))H(\mathbf{z}_i^b,\mathbf{w}_r^b) \qquad (2)$$

La fonction de coût peut être encore réécrite :

$$E(\varphi,W)=(1-F).E_{som}(\varphi,W^r)+F.E_{bin}(\varphi,W^b) \qquad (3)$$

Où

$$E_{som}(\varphi,W)=\sum_{\mathbf{z}_i\in App}\sum_{r\in C}K^T(\delta(\varphi(\mathbf{z}_i),r))\,\|\,\mathbf{z}_i^r-\mathbf{w}_r^r\,\|^2 \qquad (4)$$

est la fonction de coût classique utilisée par l'algorithme de Kohonen (la version batch), [13].

Et

$$E_{bin}(\varphi,W)=\sum_{\mathbf{z}_i\in App}\sum_{r\in C}K^T(\delta(\varphi(\mathbf{z}_i),r))H(\mathbf{z}_i^b,\mathbf{w}_r^b) \qquad (5)$$

est la fonction de coût classique utilisée par l'algorithme BinBatch (Lebbah, Badran et al. 2000). Dans le cas particulier où $F \in \{0,1\}$ la fonction de coût (3) est réduite à la fonction de coût (4) ou (5) utilisées respectivement dans le cas numérique et binaire. Pour les autres valeurs de F les deux parties sont prises en compte avec leurs pondérations. Le choix du paramètre F est déterminé par expérimentation.

Pour un paramètre F fixé, La minimisation de la nouvelle fonction de coût globale (3) est réalisée à l'aide d'une procédure itérative en deux phases:

Phase d'affectation: mise à jour de la fonction d'affectation φ associée à l'ensemble W fixé. On affecte chaque observation $\mathbf{z}$ au référent défini à partir de l'expression suivante:

$$\forall \mathbf{z}, \varphi(\mathbf{z}) = \arg\min_{c}((1-F)\,\|\,\mathbf{z}^r - \mathbf{w}_c^r\,\|^2 + FH(\mathbf{z}^b, \mathbf{w}_c^b)) \qquad (6)$$

Phase d'optimisation: La fonction d'affectation étant fixée à sa valeur courante, choisir le système de référents qui minimise la fonction $E(\varphi,W)$ dans l'espace $R^n \times \beta^m$. Ceci nous amène à minimiser la fonction $E_{som}(\varphi,W)$ (4) dans R^n et la fonction $E_{bin}(\varphi,W)$ (5) dans β^m. Ces deux minimisations permettent de définir les expressions nécessaires pour calculer l'ensemble des référents:

la partie numérique $\mathbf{w}_c^r$ du vecteur référent $\mathbf{w}_c$ est le vecteur moyen défini comme suit:

$$\mathbf{w}_c = \frac{\sum_{\mathbf{z}_i \in A} K(\delta(c, \varphi(\mathbf{z}_i)))\mathbf{z}_i^r}{\sum_{\mathbf{z}_i \in A} K(\delta(c, \varphi(\mathbf{z}_i)))},$$

la partie binaire $\mathbf{w}_c^b$ du vecteur référent $\mathbf{w}_c$ est le centre médian de la partie binaire des observations $\mathbf{z}_i \in A$ pondérées par $K(\delta(c, \varphi(\mathbf{z}_i)))$. Chaque composante $w_c^b = (w_c^{b1}, ..., w_c^{bk}, ..., w_c^m)$ est calculée comme suit:

$$w_c^{kb} = \begin{cases} 0 & \text{si} \left[\sum_{\mathbf{z}_i \in A} K(\delta(c, \varphi(\mathbf{z}_i)))(1 - \mathbf{z}_i^{bk}) \right] \geq \left[\sum_{\mathbf{z}_i \in A} K(\delta(c, \varphi(\mathbf{z}_i)))\mathbf{z}_i^{bk} \right] \\ 1 & \text{sinon} \end{cases}$$

La minimisation de la fonction de coût $E(\varphi,W)$ s'effectue par itération successive des deux phases jusqu'à stabilisation ou jusqu'à un nombre d'itérations définies à l'avance. A la fin de l'apprentissage, $\mathbf{w}_c$ partage le même codage que les observations initiales, ce qui permet une interprétation symbolique de la partie binaire du référent. La qualité de la partition résultat de la carte topologique ainsi que l'ordre topologique fourni par la grille, dépend fortement de la fonction voisinage K . Dans la pratique, comme dans le cas des cartes topologiques classiques, nous utilisons une fonction noyau avec un paramètre T pour contrôler la taille du voisinage définie par : $K^T(\delta(c,r)) = \exp(\frac{-0.5\delta(c,r)}{T})$. Ainsi, par analogie avec l'algorithme de Kohonen, les deux itérations précédentes sont répétées en faisant décroître le paramètre T entre deux valeurs T_{max} et T_{min} .

3.3 Application des approches d'apprentissage supervisé aux puces à ADN

Les approches supervisées, appelées aussi analyses discriminantes de classes ou prédictions de classes, sont des méthodes qui sont capables d'apprendre une séparation entre les groupes, suite à une phase d'apprentissage qui construit un modèle permettant la séparation entre les classes. Ce modèle est obtenu à partir d'une base pour laquelle les classes sont connues d'avance. Avant son application sur les nouvelles données, les performances de ce modèle sont estimées et les paramètres sont optimisés pour améliorer les performances du modèle. Par la suite, le modèle optimal est appliqué sur les données nouvelles.

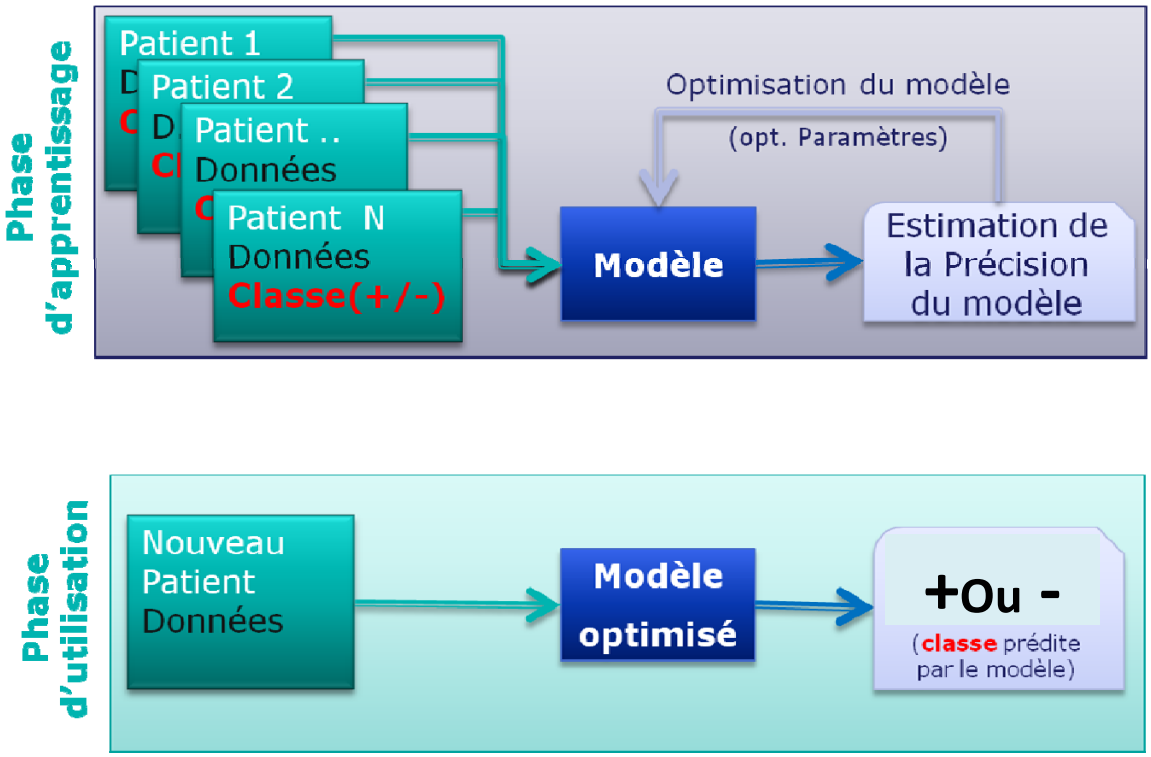


Figure 14: Principe général de la construction et de l'utilisation d'un modèle d'apprentissage supervisé

Dans ce qui suit, nous présentons les méthodes les plus utilisées dans le cadre de l'analyse des données biopuces.

3.3.1 **K plus proches voisins**

La méthode des K plus proches voisins (k-ppv) (Hart 1973)est une méthode qui consiste à rechercher dans la base d'apprentissage les individus les plus proches d'une nouvelle donnée. La règle de décision consiste à faire un vote majoritaire sur les classes de ces *k-ppv*. On peut noter qu'il est préférable de choisir une valeur de k impaire pour éviter le problème d'égalité de classe lors de la prise de décision. La méthode repose alors sur un critère de similarité qu'il faut définir a priori pour comparer les données. Le seul paramètre à régler est alors le nombre de voisins à considérer.

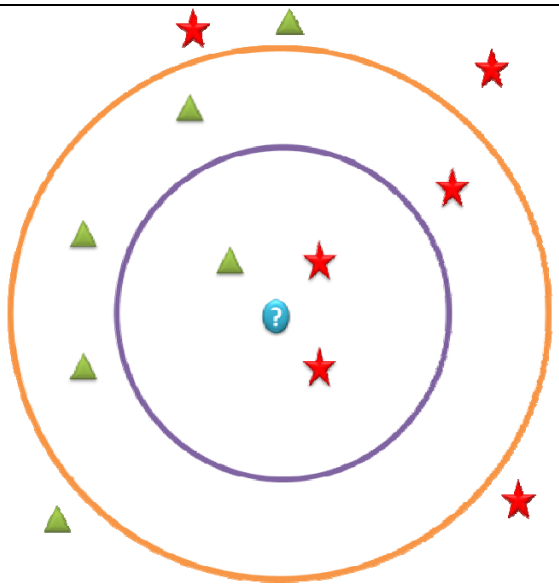


Figure 15 : exemple de classification avec les Kppv. Selon le choix du nombre de voisins, l'exemple test en bleu est classé rouge pour k=3 par contre il est classé vert pour k=7

Cette méthode peut paraître simpliste, mais dans de nombreux cas réels elle s'avère efficace, et même plus performante que des modèles plus complexes. Elle peut par conséquent, constituer une bonne référence pour quantifier les performances de classification des autres méthodes. Des études comparatives des performances de ces méthodes dans le domaine de la classification des puces à ADN des données du cancer ont été publiées par Dudoit et al. (Dudoit, Fridlyand et al. 2002). Cette méthode est employée aussi pour l'estimation des valeurs manquantes dans les puces à ADN (Troyanskaya, Cantor et al. 2001).

3.3.2 **Les méthodes d'analyse discriminante**

Dans le cadre de l'analyse des données puces, la méthode d'analyse discriminante (Fisher 1936) et ses dérivées sont des méthodes qui recherchent une combinaison linéaire des gènes qui rendent simultanément maximale la distance entre les groupes et minimales les distances entre les individus d'un même groupe.

Soient $x=(x_1,\ldots,x_n)$ les niveaux d'expression des y gènes d'un individu i, et $\{\mu_k;k=1,\ldots,m\}$ les profils génétiques moyens des m groupes. Soit $\sum_k$ la matrice de variance-covariance du groupe k.

La règle de décision de l'analyse discriminante revient à affecter un individu dont le profil génétique est décrit par x à la classe dont le profil génétique moyen est le plus proche. En faisant l'hypothèse que les niveaux d'expression des gènes suivent une loi normale multivariée $(x \mid y=k \sim N(\mu_k,\sum_k))$, ceci revient à minimiser la quantité $\left[(x-\mu_k)\sum_{-1}^{k}(x-\mu_k)' + \log\left|\sum_k\right|\right]$. Dans le cas où cette matrice diagonale est commune aux k classes, $\sum_k = \sum = diag(\sigma_1^2,\ldots,\sigma_m^2)$, on parle d'analyse discriminante diagonale linéaire(DLDA). Cette méthode a montré son efficacité dans plusieurs analyses de données biopuces (Dudoit, Fridlyand et al. 2002; Dudoit and Fridlyand 2003; Dettling and Buhlmann 2004)

3.3.3 **Les forêts aléatoires**

La méthode des forêts aléatoires (Breiman 2001) est une technique d'apprentissage supervisée qui combine une technique d'agrégation, le bagging (Breiman 1996), et une technique particulière d'induction d'arbres de décision. Cette technique a pour objectif de réduire la variabilité du prédicteur obtenu par les arbres binaires de classification (Breiman, Friedman et al. 1983), très instables, en combinant leurs résultats. La méthode CART est basée sur un découpage, par des hyperplans, de l'espace engendré par les variables. Une forêt aléatoire consiste en un nombre arbitraire d'arbres simples, utilisés pour calculer un vote pour la classe la plus représentée (classification), ou dont les réponses sont combinées pour obtenir une estimation de la variable dépendante (régression). Lors de la construction de l'arbre, pour initier la segmentation d'un nœud, la méthode effectue une sélection aléatoire parmi les variables candidates puis elle cherche la variable de segmentation. La taille de la sélection est un paramètre de l'algorithme.

La méthode de forêts aléatoires a été appliquée dans le cadre de l'analyse des biopuces (Izmirlian 2004; Svetnik, Liaw et al. 2004; Diaz-Uriarte and Alvarez de Andres 2006) et les résultats empiriques montrent l'intérêt de l'utilisation de cette approche.

3.3.4 **Les machines à vecteurs de supports (SVM)**

Les machines à vecteurs de support appartiennent à la famille des classeurs binaires (Burges, 1998, Schölkopf et al., 1998, Cristianini and Shawe-Taylor, 2000), le principe de cette méthode repose sur l'hypothèse de la recherche d'une surface optimale de séparation des deux classes qui maximise la marge. La théorie des SVM est basée sur la minimisation du risque structurel, elle vise à éviter le sur-apprentissage (over-fitting) fournissant ainsi une solution partielle au dilemme du compromis biais-variance (Vapnik 1995). Les deux éléments clés de l'implémentation de SVM sont les techniques de programmation quadratiques et les *fonctions noyaux*.

Nous nous intéresserons particulièrement à la classification binaire, les classes P+ et P- auront pour valeurs respectives de $y_i = \{+1, -1\}$. Les problèmes de classification à K-classes peuvent être obtenus par construction à partir de classeurs binaires.

L'interprétation géométrique des classeurs à vecteur de support peut être vue comme une recherche de surface de séparation optimale, un hyperplan, qui soit équidistant des frontières des deux classes (Burges 1998).

Dans ce qui suit, nous présenterons le cas où les données sont linéairement séparables, ensuite nous introduirons les fonctions noyaux afin d'expliquer la construction des surfaces de décision non linéaire. Enfin, pour les données bruitées, et dans le cas où la séparation totale des données est ambiguë, nous introduirons une variable d'écart qui autorisera une certaine marge d'erreur dans la classification.

3.3.4.1 **Machine à vecteurs de support linéaire**

Pour expliquer le fonctionnement des SVM, voici un exemple simple en deux dimensions dont les classes sont linéairement séparables, une droite sépare les classes « P+ » des classes « P-».

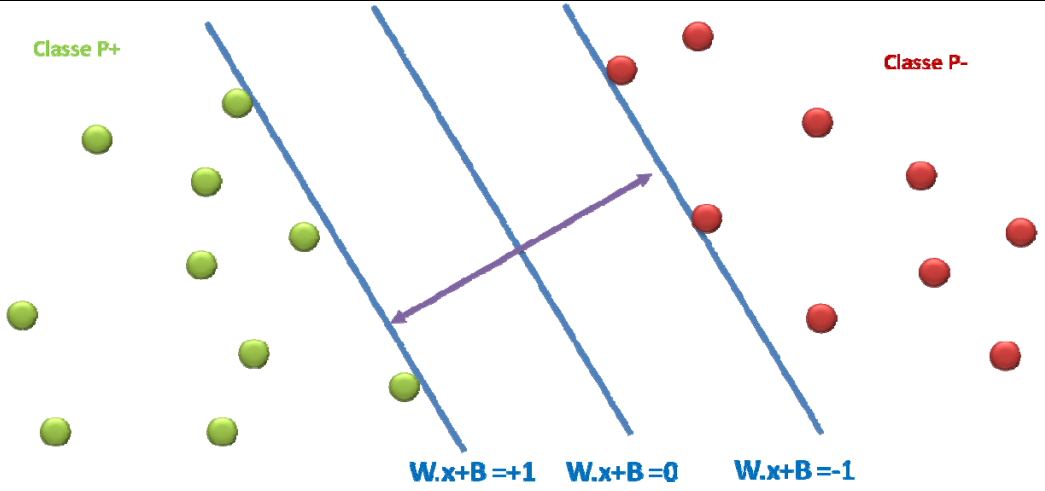


Figure 16 : Problème de classification linéairement séparable

Dans la Figure 16, la surface de séparation est définie par :

$$\mathbf{w}\cdot\mathbf{x}+b=0 \qquad (1)$$

Le but est donc de trouver un vecteur **w** et un scalaire *b* qui assurent une bonne classification pour chaque point. Autrement dit, on cherche **w** et *b* qui satisferaient l'inégalité suivante :

$$y_i(\mathbf{w}\cdot\mathbf{x_i}+b)\geq 1,\ \forall i\in[1..m] \quad (2)$$

On construit les deux plans de frontières d'équations $\mathbf{w}\cdot\mathbf{x_i}+b=1$ et $\mathbf{w}\cdot\mathbf{x_i}+b=-1$

Tout couple $(\mathbf{w},b)$ satisfaisant la contrainte (2) séparera d'une façon appropriée les points des deux classes. La tâche suivante est de trouver le meilleur choix de $(\mathbf{w},b)$ qui permettra non

seulement de bien classer les données d'apprentissage mais aussi les données inconnues au classeur. Pour trouver le meilleur plan de séparation, on cherche des plans frontières qui soient éloignés l'un de l'autre tout en maintenant une bonne précision de classification. La distance de séparation entre les deux plans frontières est égale à $\frac{2}{\sqrt{\mathrm{w} \times \mathrm{w}}}$. Pour maximiser cette distance, on cherche à minimiser **w e**n respectant la contrainte (2):

$$\min_{\mathbf{w},b} \frac{1}{2}\mathbf{w} \cdot \mathbf{w} \qquad (3)$$
$$\text{Sous contrainte } y_i(\mathbf{w} \cdot \mathbf{x_i} + b)_i \geq 1, \forall\ i \in [1..m]$$

Ce type de problème d'optimisation est connu sous le nom de programme quadratique, il est caractérisé par une expression quadratique à minimiser et des contraintes linéaires.

On considère maintenant les problèmes ou les classes ne sont pas linéairement séparables, dans ce cas de figure on cherche à trouver le meilleur couple $(\mathbf{w},b)$ qui classe au mieux les exemples. Pour cela, on introduit alors une variable d'écart ξ dans la contrainte(2), ξ est non nulle seulement pour les points qui sont mal classés. Le problème s'écrit alors :

$$\min_{\mathbf{w},b,\xi} \frac{1}{2}\mathbf{w} \cdot \mathbf{w} + C\sum_{i=1}^{m} \xi_i, \text{Sous contrainte } y_i(\mathbf{w} \cdot \mathbf{x_i} + b) + \xi_i \geq 1, \forall\ i = 1,2,...,m \qquad (4)$$

La fonction objectif du programme quadratique(4) contient deux termes. Le terme $\mathbf{w} \cdot \mathbf{w}$ vise à maximiser la distance entre les frontières de séparation et le terme $\sum_{i=1}^{m} \xi_i$ réduit l'erreur en classification. Le paramètre positif C est introduit pour calibrer le résultat obtenu. Le choix d'une grande valeur de C précise que plus d'importance est accordée à la réduction de l'erreur en classification. Inversement, une petite valeur de C assure une bonne séparation des surfaces frontières permettant ainsi d'éviter le sur-apprentissage (overfitting). La bonne valeur de C, est retrouvée expérimentalement en effectuant des tests via la méthode de la validation croisée. Des techniques plus sophistiquées pour retrouver C ont été proposées (Wahba, Lin et al. 2000).

Les points se situant sur les frontières ainsi que les points qui sont au-delà des bonnes frontières sont appelés les vecteurs de support (voir figure 2). Ces vecteurs de support sont très importants, en effet, si on garde ces vecteurs de support et que l'on retire tous les autres points, le problème d'optimisation fournit la même solution.

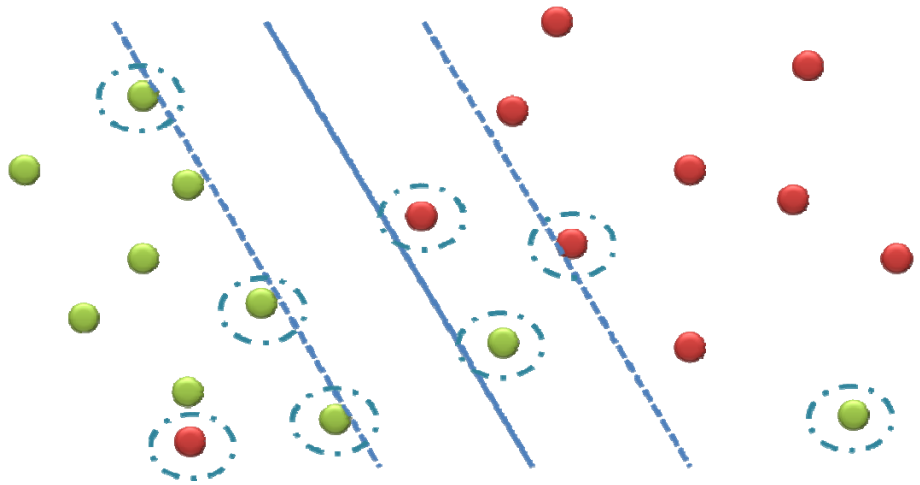


Figure 17 : Exemple de classification en deux dimensions avec vecteur de support indiqué par des cercles

3.3.4.2 **Machine à vecteurs de supports non linéaire**

Ce cas d'utilisation des machines à vecteurs de support est intéressant car la plupart des problèmes réels sont non linéairement séparables.

Tout programme quadratique soluble possède un équivalent dual qui est plus facile à calculer. En introduisant le Lagrangien, le problème dual équivalent au programme quadratique (4) s'écrit (Mangasarian 1994; Schölkopf, Burges et al. 1998; Cristianini and Shawe-Taylor 2000) :

$$\max_{\boldsymbol{\alpha}} \frac{1}{2}\sum_{i,j=1}^{m} y_i y_j \alpha_i \alpha_j K(\mathbf{x}_i \cdot \mathbf{x}_j) - \sum_{i=1}^{m} \alpha_i$$
$$\text{sous containte} \begin{cases} \sum_{i=1}^{m} y_i \alpha_i = 0 \\ 0 \le \alpha_i \le C,\ i \in [1..m] \end{cases} \qquad (5)$$

Le vecteur α est la variable duale, elle remplace $(\mathbf{w},b)$ dans la formulation primale. Cette formulation duale peut être généralisée afin de trouver les surfaces de séparation non linéaire. Dans cette formulation, on se sert plutôt du produit scalaire $x_i.x_j$ entre les points que des points eux même. Par conséquent, on peut se servir de l'astuce du noyau (kernel trick) pour remplacer le produit scalaire par une fonction noyau. Ces fonctions sont des fonctions non linéaires et elles jouent un rôle similaire au rôle du produit scalaire dans les problèmes d'optimisation. On utilisera souvent la version matricielle de ce problème. Pour ce faire, on introduit la matrice **G** telle que $G_{ij} = y_i k(x_i, x_j) y_j$. Cette matrice G, appelée Hessienne[3], est donc très proche de la matrice de Gram[1] et possède les propriétés de symétrie et de définition positive. Le problème d'optimalisation (5) peut s'écrire alors sous la forme matricielle suivante :

$$\max_{\alpha} \frac{1}{2}\alpha^T G\alpha - 1^T\alpha$$
$$\text{sous containte} \begin{cases} y^T\alpha = 0 \\ 0 \le \alpha \le C, \forall i \in [1,n] \end{cases} \quad (6)$$

Les problèmes d'optimisation (5) et (6) sont des problèmes d'optimisation convexes. Toute solution locale est aussi une solution globale, cette propriété est due au fait que la matrice du noyau soit définie positive. Par conséquent, en ayant trouvé une solution, on est sûr qu'il s'agit bien d'une solution optimale.

3.3.4.3 **Fonctions noyaux**

D'une manière générale, il peut être utile de savoir à quel point un exemple est similaire à un autre. On utilise souvent en mathématique le produit scalaire qui, moyennant une normalisation, correspond au cosinus de l'angle entre deux vecteurs, donc une fonction noyau peut être vue comme une mesure de similarité.

[3] http://www.ee.ic.ac.uk/hp/staff/dmb/matrix/

Soit Φ une application non linéaire de $\aleph$dans $\Im$, où $\aleph$ est un espace de données initiales, et $\Im$ un espace des descripteurs (en général de dimension supérieure), un noyau est défini comme suit :

$$K(x,z)=\langle \Phi(x),\Phi(z)\rangle \qquad (7)$$

Une propriété importante est la condition de Mercer (Mercer 1909) :

La fonction $K(x,z):(\aleph,\aleph)\to R$ est un noyau si et seulement si $G(k(x_i,x_j))_{i,j=1}^{n}$ est définie positive.

Voici quelques exemples de noyaux usuels utilisés dans la littérature (Burges 1998; Cristianini and Shawe-Taylor 2000) :

- Noyau linéaire : $K(\mathbf{x_i},\mathbf{x_j})=(\mathbf{x_i}\cdot\mathbf{x_j})$
- Noyau Polynomial : $K(\mathbf{x_i},\mathbf{x_j})=(\mathbf{x_i}\cdot\mathbf{x_j}+1)^d, d\in R_+^*$
- Noyau Gaussien (Radial): $K(\mathbf{x_i},\mathbf{x_j})=e^{-(\mathbf{x_i}-\mathbf{x_j})\cdot(\mathbf{x_i}-\mathbf{x_j})/2\sigma^2}, \sigma\in R_+^*$

La matrice contenant les mesures de similarité entre tous les exemples de l'ensemble d'apprentissage est une matrice carrée définie positive, elle est appelée matrice de Gram,

$$\begin{pmatrix} k_{11} & \cdots & k_{1n} \\ \vdots & \ddots & \vdots \\ k_{n1} & \cdots & a_{nn} \end{pmatrix}$$

Il est possible de composer de nouveaux noyaux à partir de noyaux. Soit K_1 et K_2 des fonctions satisfaisant la condition de Mercer et B une matrice définie positive, alors les fonctions suivantes sont des noyaux.

$$K(x,z)=K_1(x,z)+K_2(x,z)$$
$$K(x,z)=\mu.K_1(x,z), \mu\in R$$
$$K(x,z)=K_1(x,z)*K_2(x,z)$$
$$K(x,z)=x^TBz$$

Étant donné un noyau K, pour prédire la classe d'une nouvelle donnée on calcule le signe de la fonction *f* définie par :

$$signe\left(f(\mathbf{x}) = \sum_{i=1}^{m} y_i \alpha_i^* K(\mathbf{x}, \mathbf{x_i}) + b^* \right) \quad (8)$$

α^* est la solution du problème quadratique dual (8) , b^* vérifie l'équation $y_i f(x_i) = 1, \forall i$ avec $0 < \alpha_i < C$, K une fonction noyau (Cristianini and Shawe-Taylor 2000). Si le résultat obtenu est positif la nouvelle entrée appartient à la classe « P+ » sinon elle appartient à la classe « P-».

3.4 Estimation des performances d'un modèle

Apres avoir vu les différents modèles, nous présentons dans ce qui suit les méthodes qui permettes d'évaluer les performances de ces modèles.

L'estimation des performances d'un modèle se mesure en termes de précision de la prédiction sur la base de test qui n'a pas servi pour la construction du modèle.

L'idée est de disposer d'un ensemble de mesures de qualité permettant de tester la qualité de la procédure de classification. Pour ce faire, on partitionne l'échantillon en un *ensemble d'apprentissage* et un *ensemble test*. La répartition entre les deux ensembles doit être faite d'une façon aléatoire. L'estimation de l'erreur réelle est alors l'erreur apparente mesurée sur l'ensemble test. La qualité de l'apprentissage augmente avec la taille de l'ensemble d'apprentissage, de même, la précision de l'estimation augmente avec la taille de l'ensemble test. Mais, dans la pratique, la taille de l'échantillon est limitée. Cette méthode donne de bons résultats lorsque l'échantillon est 'assez' grand. Il existe peu de résultats théoriques sur les tailles d'échantillons nécessaires pour utiliser cette méthode, on ne dispose que de résultats empiriques qui dépendent du problème. La répartition de l'échantillon entre les deux ensembles se fait en général dans des proportions 2/3 pour l'ensemble d'apprentissage et 1/3 pour l'ensemble test. Pour les échantillons de petite taille, deux méthodes sont fréquemment utilisées, la *validation croisée* et la méthode du *bootstrap (Efron 1993)*. La *validation croisée*

consiste à découper l'échantillon en *k* sous-ensembles. Un ensemble d'apprentissage consiste en la réunion de *k*-1 sous-ensembles et un ensemble test au *k*-ième sous ensemble. On exécute alors l'apprentissage sur chacun des *k* ensembles d'apprentissage et on estime l'erreur réelle par l'erreur apparente sur l'ensemble test correspondant. L'estimation de l'erreur réelle est alors la moyenne des erreurs apparentes obtenues. La méthode du *bootstrap* consiste à construire, à partir de l'échantillon *S* de tailles *n, m* ensembles d'apprentissage de taille *n* (un élément de *S* peut ne pas appartenir à l'ensemble d'apprentissage, ou y figurer plusieurs fois). L'ensemble *S* sera utilisé comme un ensemble test. L'estimation de l'erreur réelle est alors la moyenne des erreurs apparentes obtenues pour un certain nombre d'itérations de l'algorithme d'apprentissage. Ces deux méthodes fournissent de bons estimateurs de l'erreur réelle mais sont très coûteuses en temps de calcul.

Chapitre 4

Prédiction de la perte

de poids chez les patients obèses

4.1 Le cadre du projet NUGENOB.

4.1.1 **Introduction**

La complexité biologique de l'obésité donne à penser que l'utilisation des puces à ADN pour identifier des gènes capables d'élucider le fonctionnement du métabolisme du tissu adipeux et son altération au cours de l'obésité est une approche prometteuse. Ceci est principalement dû au fait qu'il existe une grande variabilité interindividuelle quant à la réponse à une intervention diététique, variabilité dont l'origine demeure principalement inconnue(Viguerie, Poitou et al. 2005).En effet, l'obésité n'est pas la conséquence de la défaillance d'un seul gène mais plutôt la conséquence d'interactions gène-gène et gène-environnement qui varie d'un individu à un autre (Klaus and Keijer 2004; M. J. Moreno-Aliaga 2005; Mutch and Clément 2006). Alors que la plupart des travaux sur les puces à ADN dans le cadre de l'obésité ont mis l'accent sur la comparaison des catégories (comparaison des obèses versus maigres, préadipocytes versus adipocytes, etc), la différence entre la notion de comparaison de classe et celle de la prédiction est souvent floue et les résultats sont souvent orientés vers la recherche de nouveaux marqueurs génétiques pour faire le pronostic de la maladie (Ein-Dor, Kela et al. 2005; Lin, Devakumar et al. 2006).

Pourtant, il ya une limitation inhérente à une telle approche puisque les modèles ont tendance à 'sur-apprendre' les données (Perez-Diez, Morgun et al. 2007). Avant de prétendre qu'un gène

ou un ensemble de gènes est considéré comme un bon prédicteur, une validation de l'hypothèse doit être effectuée sur un des échantillons indépendants. Car dans la plupart des cas un prédicteur donne de bons résultats sur les échantillons utilisés pour l'identifier mais cela ne prouve pas qu'il sera performant sur des données qui lui sont 'inconnues'. Ce problème peut être évité en utilisant les approches d'apprentissage supervisé et les méthodes d'estimation et ainsi fournir un modèle prédictif fiable. Dans le cadre du projet européen NUGENOB, nous nous sommes intéressés à la question de savoir si l'expression des gènes du tissu adipeux sous-cutané est capable de prédire la perte de poids d'un individu ayant suivi un régime hypocalorique faible en matière grasse. Le choix du tissu adipeux sous cutané pour la réalisation des puces à ADN est motivé par facilité et la rapidité des ponctions pratiqués à l'aiguille.
Même si on sait peu de choses en ce qui concerne les changements dans l'expression des gènes adipeux encourus à la suite de la restriction d'énergie à faible teneur en matières grasses et riche en glucides (Viguerie, Vidal et al. 2005; Sorensen, Boutin et al. 2006), on en sait encore moins sur la capacité du profil de l'expression génique du tissu adipeux humain à prédire la réponse en terme de perte de poids suite à un régime alimentaire.

Le but de la présente analyse est d'étudier si l'expression des gènes avant les 10 semaines de régime alimentaire hypocalorique pourrait être utilisée pour prédire avec précision la réponse d'une personne obèse à un régime de ce type. Par le biais de l'application d'une combinaison d'approches statistiques et de techniques d'apprentissage supervisé, nous démontrons que la possibilité de distinguer les profils d'expression des intervenants dans le régime ne garantit pas une grande précision pour ce qui est de la prédiction, cela en contraste avec le succès de la même tâche réalisée sur des données publiques sur le cancer (Golub, Slonim et al. 1999).

La présente étude a porté sur 54 femmes recrutées dans le carde du projet NUGENOB, ayant suivi un régime hypocalorique faible en matière grasse (lipides de l'alimentation) de 10 semaines. Les tests statistiques ont révélé que les profils d'expression génique des sujets répondeurs (8-12 kg perte de poids) peuvent toujours être dissociés de non-répondeurs (<4 kg de perte de poids). Nous avons également évalué si cette différenciation est suffisante pour la

prédiction. Les approches d'apprentissage supervisé fournissent des modèles de prédiction avec une précision maximale de 61,1% ± 8,1%.

4.1.2 **Sélection des sujets pour l'analyse prédictive**

Dans le cadre de ce projet européen, 771 sujets ont suivi un des deux régimes hypocaloriques mis en place pour les études envisagées : un régime à faible teneur en matières grasses et riche en glucides (LF), ou un régime modéré en matières grasses et faible en glucides (MF). La présente étude ne concerne que les femmes du groupe(LF). Le régimeLF(conçu pour fournir une diminution de 600 kcal / jour des apports habituels)a été suivi pendant 10 semaines. Des biopsies du tissu adipeux sous-cutané ont été obtenues pour la majorité des 771 sujets participant à cette étude d'intervention diététique (à la fois avant et après l'intervention diététique). Un spécimen de graisse sous-cutanée abdominale (~ 1 g) a été obtenu par aspiration sous anesthésie locale. Les biopsies ont été stockées à -80 ° C jusqu'au moment de l'analyse. Après élimination des sujets ayant abandonné l'étude et ceux dont la quantité d'ARN extraite était insuffisante, 319 sujets ont été évalués pour la perte de poids après 10 semaines et ensuite divisé en deux groupes: les «répondeurs» (sujets ayant perdu entre 8 et 12 kg) et les «non-répondeurs» (sujets ayant perdu moins de 4 kg). Vingt-sept sujets de sexe féminin ont été choisis au hasard de chaque groupe après une correspondance en fonction du poids, taille, indice de masse corporelle (IMC), rapport de tour de taille / hanche, l'apport énergétique, les lipides, les glucides et les protéines. Seuls les ARN totaux à partir de biopsies prises avant les 10 semaines d'intervention alimentaire ont été utilisées dans la présente étude prédictive.

4.1.3 **Données Leucémie**

Pour évaluer les performances des modèles, une base de données de puce de type Affymetrix portant sur 47 sujets avec la leucémie aiguë lymphoïde (ALL) et 25 sujets de leucémie aiguë myéloïde (AML), comme décrite précédemment par Golub et ses collègues (Golub, Slonim et al. 1999), a été téléchargée à partir du lien suivant : http://wwwmaths.anu.edu.au/johnm/r/hddplot/. Le jeu de données de la base du cancer

“Golub” contient l’expression de 7129 gènes communs à tous les 72 sujets de l’étude. Les puces à ADN ont été normalisées par la médiane.

4.1.4 **Analyse prédictive à partir des données biopuces**

54 puces à ADN, correspondant à 27 puces de sujets « répondeurs » et 27 puces de sujets « non-répondeurs », ont été réalisées. Cependant, une puce d’un sujet non-répondeur était de mauvaise qualité et a été éliminée de l’analyse. Les 53 puces à ADN restantes ont été normalisées par la méthode loess (Yang, Dudoit et al. 2002). 14135 gènes étaient présents dans 80% des puces à ADN, représentant 76,4% du total des gènes sur les puces 44K. L'expression différentielle de gène, en utilisant 5% de taux de fausses découvertes (FDR),a été réalisée avec l’outil SAM (Significance Analysis of Microarrays), disponible à la fois en implémentations Excel et R (Tusher, Tibshirani et al. 2001).Un fisher et un t-test ont également été réalisés. Les trois approches statistiques ont été utilisées en raison de récents rapports démontrant que les différents algorithmes peuvent générer des listes de gènes avec différents degrés de pouvoir prédictif (Ein-Dor, Kela et al. 2005; Lin, Devakumar et al. 2006). Pour l’analyse prédictive, seuls les 10592 gènes pour lesquels il n'existe pas de valeurs manquantes dans toutes les puces à ADN ont été retenus.

Pour l’analyse prédictive, nous avons utilisé les méthodes de référence dans le domaine de la prédiction à partir des puces à ADN : les K plus proches voisins (Deegalla and Boström 2007), les méthodes d’analyse discriminante (Dudoit, Fridlyand et al. 2002; Diaz-Uriarte and Alvarez de Andres 2006), les forêts aléatoires (Liaw and Wiener 2002; Diaz-Uriarte and Alvarez de Andres 2006) et les machines à vecteurs de supports (Furey, Cristianini et al. 2000; Dudoit, Fridlyand et al. 2002; Diaz-Uriarte and Alvarez de Andres 2006).

Tous les algorithmes sont développés sous R: KNN est implémenté dans le package ensemble, DLDA dans le package sma, RF est dans le package randomForest et SVM dans le package Kernlab. Les paramètres standards ont été utilisés pour toutes les méthodes sauf pour SVM, où le les paramètres ont été optimisés en utilisant le package svmpath (Trevor Hastie 2004).

4.1.4.1 **Résultats**

Des puces à ADN ont été réalisées à partir de l'ARN d'un tissue adipeux prélevé avant le régime faible en matière grasse et riche en glucides de 10 semaines afin d'étudier la faisabilité de la mise en place d'un système d'aide à la décision capable de prédire la perte de poids suite à un régime à partir des données initiale récupérées avant l'intervention diététique.

4.1.4.2 **Différentiation entres les répondeurs et les non-répondeurs**

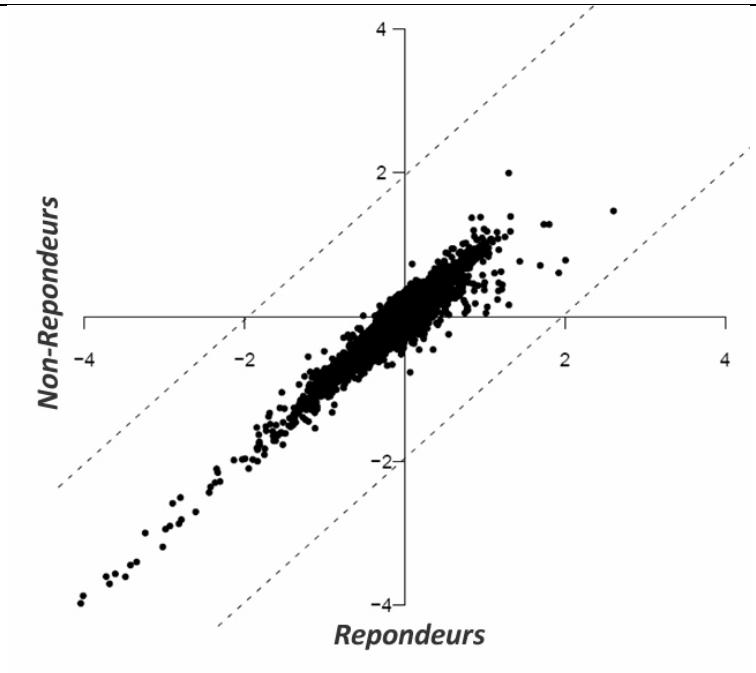


Figure 18 : Distribution de la moyenne des niveaux d'expression de gènes normalisés chez les répondeurs et les non-répondeurs. Chaque spot représente la moyenne pour l'expression d'un seul gène. Les lignes en pointillé indiquent l'intervalle de confiance à 95%.

Comme l'illustre la Figure 18, l'expression génique globale était similaire entre les répondeurs et les non-répondeurs. Néanmoins, l'utilisation de 3 différentes approches statistiques nous a permis d'identifier des gènes différentiellement exprimés entre les deux populations. Une analyse SAM avec une FDR de 5% n'a pas révélé de différences entre les deux

populations par contre avec une contrainte moindre sur le FDR passée à 8% l'analyse a permis d'identifié 34 gènes différentiellement exprimés. Les critères de sélection pour les deux autres tests analytiques (Fisher et test de Student) ont été arbitrairement fixés, nous avons fixé les 100 meilleurs gènes qui différenciaient les deux populations. Afin de réduire le biais qui pourrait être induit par la sensibilité des tests statistiques, nous avons considéré les gènes résultants de l'intersection des 3 listes générées par chacun des tests.

Neuf gènes ont été identifiés comme étant sensiblement augmentés chez les non-répondeurs par rapport aux répondeurs: la prostaglandine D2 synthase (Ptgds: 1,6 fois), Claudin 5 (Cldn5: 1,4 fois), fibromodulin (FMOD: 1,4 fois), l'interféron alpha-inducible protein 27 (Ifi27: 1,4 fois), quinolinate phosphoribosyltransférase (Qprt: 1,3 fois), de la famille avec la similitude de séquence 69 B (Fam69b: 1,3 fois), protéines transmembranaires 132A (Tmem132A: 1,2 fois), des cellules endothéliales molécule d'adhésion (Esam: 1,2 fois), et TPTE / PTEN homologue de lipides inositol phosphatase pseudogene (LOC374491: 1,1 fois). La signification statistique pour Ptgds, Cldn5, Qprt, et Tmem132A a été confirmée parRT-PCR temps réel ($p < 0,05$; tableau 2), et alors que la direction de concordance a été réalisée pour FMOD, Ifi27, Fam69b et Esam, les changements dans l'expression sont modestes. Néanmoins, ces 9 gènes ont été considérés comme les candidats aptes à évaluer leur capacité à prédire la perte de poids.

Table 12: Validation des 8 prédicteurs par RT-PCR temps réel

Nom gene	FC microarray	RT-PCR quantitative en temps réel	P-value RT-PCR
TMEM132A	1.2	2.6	0.008
QPRT	1.3	2.1	0.015
CLDN5	1.4	2.3	0.015
PTGDS	1.6	1.9	0.035
ESAM	1.2	2.2	0.126
FMOD	1.4	2.3	0.159
FAM698	1.3	2.1	0.176
IFI27	1.4	1.4	0.182

Afin de déterminer si le profil d'expression génique pourrait permettre de distinguer les répondeurs des non-répondeurs, une analyse discriminante (PLS-DA) a été réalisée.

La Figure 19.A montre que les profils des deux groupes se séparent en utilisant l'expression génique des tissus adipeux sous-cutanés, cependant, il existe un certain chevauchement entre les deux populations (R^2 = 0,547 et Q^2 = -0,096, où R^2 explique la variation cumulative des deux premières composantes et Q^2 indique la variation du modèle en fonction de la validation croisée). Contrairement aux résultats obtenus à partir de l'expression génique du tissu adipeux, la Figure 19.B montre clairement que les données du cancer peuvent, par le moyen d'une analyse discriminante de type PLS-DA, différentier les deux populations à l'exception d'une seule aberration, comme expliqué ci-dessous (R2 = 0,795 et Q2 = 0,622).

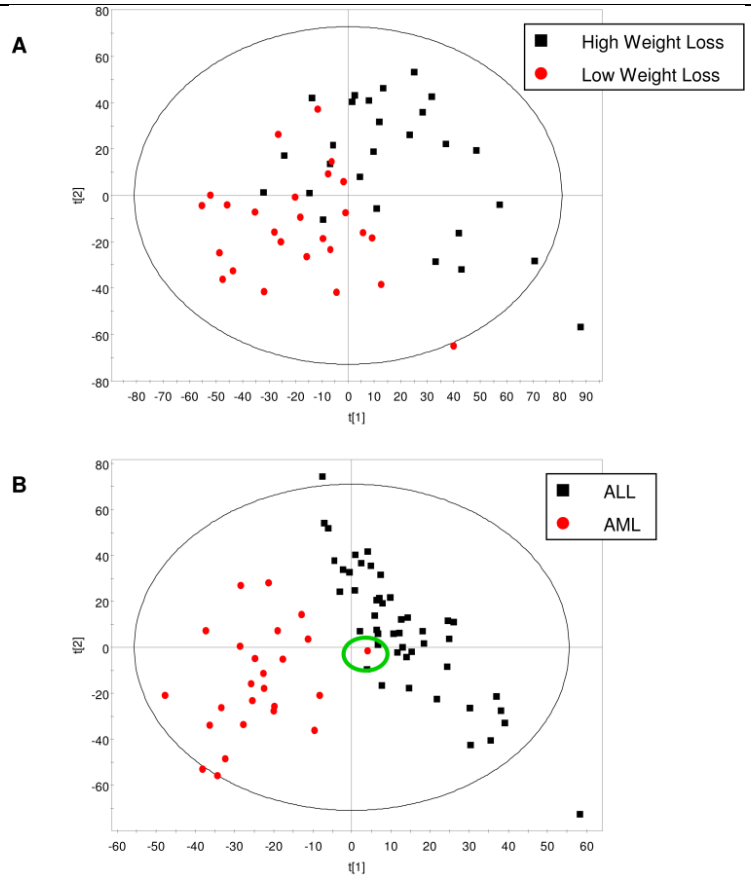


Figure 19 : différentiation des populations par une analyse discriminante (PLS-DA). Sur les données de l'obésité (A) et les données du cancer (B)

Bien que les deux premières composantes principales à elles seules soient capables de différentier les deux groupes, seul le modèle obtenu à partir des données cancer (Golub) a permis de donner des résultats de prédiction significatifs lors de la phase de validation du modèle.

4.1.4.3 **Prédiction de la perte de poids suite au régime de 10 semaines.**

Pour répondre à la question de la prédiction, nous avons utilisé une panoplie de méthodes d'apprentissage automatique supervisé de l'état de l'art. Le problème est, rappelons le, de prédire la perte de poids suite à un régime et trouver un modèle capable de prédire les répondeurs des non-répondeurs.

Pour cette analyse nous avons utilisé les machines à vecteurs de supports(SVM), les forêts aléatoires(RF), les K-plus proches voisins(KNN) et l'analyse discriminante diagonale linéaire (DLDA). La validation est faite par validation croisée (10 fois 10-folds); un modèle d'apprentissage pour chaque classeur est entrainé en utilisant 9/10^{ième} des données et le 1/10^{ième} restant est utilisé pour la phase de test (Kohavi 1995).Afin d'avoir un modèle plus robuste, nous avons lancé la validation croisée 10 fois et pris la moyenne des résultats obtenus. La précision de la prédiction des modèles a été comparée à l'approche naïve qui place tous les sujets dans la classe majoritaire (pour les données de l'obésité, la précision de la prédiction de l'approche naïve est de 51% (pour la base golub elle est de 65%), un résultat significativement supérieur à celui de l'approche naïve a été considéré comme une amélioration de la précision de prédiction. En utilisant l'intégralité de l'expression des 10592 gènes, nous obtenons une très faible amélioration de la précision de la prédiction par rapport à l'approche de référence comme l'indique la Table 13. Avec la base de l'obésité, le meilleur résultat est obtenu par la méthode KNN avec une précision de la prédiction 61,1% ± 8,1%. SVM, RF, et DLDA ont prédit la perte de poids avec une précision de 45,9% ± 5,7%, 52,9% ± 11,3% et 52,6% ± 8,2%, respectivement. En revanche, en utilisant la base de données du cancer avec les mêmes méthodes déployées précédemment, on obtient 98,6% ± 0,0%, 96,9% ± 2,2%, 92,0% ± 2,9%, et 89,0% ± 4,1% en utilisant respectivement SVM, RF, KNN, et DLDA.

Table 13 : résultat de la précision de la prédiction avec les données de l'obésité et du cancer

Données utilisées	**Méthodes d'apprentissage**			
	SVM	RF	KNN	DLDA
Données Obésité (10592gènes)	**45,9% ± 5,7%**	**52,9% ± 11,3%**	**61,1% ± 8,1%**	**52,6% ± 8,2%**
Données Cancer (7129 gènes)	**98,6% ± 0,0%**	**96,9% ± 2,2%**	**92,0% ± 2,9%**	**89,0% ± 4,1%**

Nous avons aussi tenté d'améliorer la prédiction en effectuant une sélection d'attributs. Nous avons pris les 100 meilleurs gènes en validation croisée par la méthode de Fisher ou un test-t, mais les résultats obtenus par ces deux méthodes de sélections n'ont pas abouti à une amélioration de la prédiction comme l'indique la Table 14.

Table 14 : résultat de la précision de la prédiction avec les données de l'obésité et du cancer

Données utilisées	DLDA	KNN	RF	SVM
Meilleurs 100 par Fisher-score	54.20%	53.00%	53.40%	49.30%
Meilleurs 100 par T-Test	56.60%	47.90%	51.60%	45.70%

4.1.5 **Discussion**

Réussir à prédire la réponse en termes de perte de poids d'un patient suite à un régime a un impact important d'un point de vu clinique. En effet, si un clinicien pouvait, à priori, connaitre si l'état d'un patient était susceptible de s'améliorer suite à un régime alors cela influencerait la prise en charge de la maladie sur plusieurs fronts. L'analyse de l'expression des gènes a été privilégiée dans notre étude en grande partie en raison du succès de cette approche dans le domaine de l'oncologie. La présente étude a révélé que si un aperçu de l'expression des gènes avant régime alimentaire hypocalorique de 10 semaines peut différencier les non-répondeurs des répondeurs, ces données sont insuffisantes pour la prédiction de l'effet de ce type de régime pour de futurs patients dans le cadre d'un usage en clinique.

Des études dans le domaine de l'oncologie suggèrent que la classification des tumeurs par des analyses de puces à ADN est une approche intéressante pour répondre à ces problèmes, néanmoins, malgré ces résultats encourageants, il n'existe pas actuellement de travaux montrant que cette approche peut être utile dans l'étude des maladies liées à la nutrition comme l'obésité, si l'on considère que le métabolisme du tissu adipeux est régi non seulement par la composante génétique d'un individu, mais aussi par des facteurs environnementaux (par exemple alimentation, activité physique, virus, etc.)(Petersen, Taylor et al. 2005; Mutch and Clément 2006). Comme indiqué précédemment, les différences individuelles en réponse à des interventions alimentaires sont contrôlées à la fois par la génétique et parle mode de vie (Perusse and Bouchard 2000; Loos and Rankinen 2005; Viguerie, Vidal et al. 2005).

L’analyse du profil d’expression des gènes chez 53 sujets humains ayant participé à un régime alimentaire hypo-calorique de 10 semaines dans le cadre de l’étude NUGENOB a révélé qu’il est possible de différencier les sujets répondeurs(perte de poids de 8-12 kg) des sujets non-répondeurs (perte inférieure à 4 kg), mais cette différence n’est pas aussi claire que ce qui est observé dans les données de cancer de Golub. Le meilleur résultat obtenu était de 61.1%, ces faibles performances en prédiction donnent à penser que pour cette analyse il n'y a pas de prédicateurs géniques capables de prédire l’efficacité du régime suivi, à la différence de l’étude Golub où nous avons observé 98,6% de précision. Ceci met également en évidence une différence importante quant à la question biologique, nous avons tenté de prédire la réponse à un traitement plutôt que de classer un individu selon l'état de la maladie. En tant que telle, la présente étude montre que la classification d’un type de tumeur peut-être d’une certaine façon moins difficile que la prédiction d’une réponse à un traitement.

Des exemples existent dans lesquels un gène ou un sous-ensemble de gènes ont été proposés, mais ces études concernent la comparaison entre les classes plutôt qu’une prédiction de classes. Par exemple, Tseng et al. ont effectué des analyses de puce ADN de préadipocytes et déduisent que les substrats du récepteur insuline et la necdin sont des gènes importants dans la différenciation adipocytaire (Tseng, Butte et al. 2005). Et plus récemment, Koza et al. ont analysé le profil d'expression des gènes du tissu adipeux et ont révélé que ce profil permet de distinguer des souris à faible taux de graisse de celle à taux de graisse élevé bien avant la consommation d'un régime alimentaire riche en graisses (Koza, Nikonova et al. 2006). Bien que ces études suggèrent que l'identification des gènes prédictifs à l'aide de l'expression génétique est possible, il existe plusieurs différences entre nos travaux et ceux des deux études susmentionnées. Tout d'abord, les études présentées ne sont pas des études réalisées sur des humains. Deuxièmement, aucune des études mentionnées n’utilise des approches d’apprentissage supervisé pour identifier les prédicteurs, mais des approches statistiques ont été utilisées pour identifier les gènes différentiellement exprimés. De ce fait, la valeur prédictive de leurs candidats n'a pas été évaluée. Si les études précédentes fournissent un aperçu moléculaire sur l’expansion de la masse grasse, ils leur restent certainement à démontrer que leurs cibles moléculaires sont cliniquement fiables.

A notre connaissance, l'usage de l'expression des gènes pour prédire une réponse à l'intervention diététique chez l'homme n'a pas été effectué auparavant, cependant il existe des études dans lesquelles l'efficacité de certains médicaments de l'obésité (tel que le sibutramine et l'orlistat) ont été explorées. L'efficacité de ces médicaments a été déterminée en évaluant la perte de poids dans les 3 premiers mois de traitement (Rissanen, Lean et al. 2003; N. Finer 2006). Les sujets qui ont suivi le traitement médicamenteux ont perdu du poids dans les premiers mois du traitement. Ces analyses de profils de perte sur une période de temps suggèrent que la prévision de l'évolution de la valeur de l'expression sur une période pourrait être plus informative que la valeur de l'expression juste avant le début du régime. Une telle approche permettrait d'évaluer la réponse du système biologique.

En conclusion, les puces à ADN permettent d'avoir une bonne idée sur les mécanismes biologiques qui régissent le métabolisme des tissus adipeux, mais leur usage dans un contexte clinique pour aider à la prise en charge des patients et le choix des traitements reste à un stade précoce. Néanmoins, d'autres stratégies et d'autres approches visant à maximiser la richesse des informations des puces à ADN sont à l'étude et donnerons espoir d'un système d'aide à la décision plus exploitable en clinique.

Dans la section suivante, nous parlerons du projet européen 'Diogènes' qui est la continuation du projet 'Nugenob' et des analyse que nous avons réalisé dans le cadre de ce projet.

4.2 Le cadre du projet DIOGENES.

4.2.1 **Présentation du projet Diogenes**

L'augmentation rapide de la prévalence de l'obésité et les co-morbidités est un important problème de santé mondial. En Europe, l'obésité consomme environ 5% du total des budgets de soins de santé. L'augmentation rapide de l'obésité chez les enfants de tous les pays européens nécessite des mesures d'urgence immédiates qui ont été demandées par le Conseil européen des ministres en 2002. Alors que la susceptibilité à l'obésité est largement déterminée par des

facteurs génétiques, l'actuelle épidémie d'obésité est fortement influencée par des facteurs défavorables liés au mode de vie. Compte tenu de notre bagage génétique, il est quasiment irréalisable pour l'homme d'autoréguler la prise alimentaire en vertu des conditions environnementales. Cette tendance inquiétante devrait convaincre la communauté scientifique de développer ses efforts de recherche en utilisant des approches novatrices. Le projet DiOGenes a pour objectif d'être un tel effort d'innovation, en incluant diverses disciplines qui peuvent contribuer à mieux comprendre cette maladie et trouver des moyens de prévention et de traitements diététiques.

Le projet DiOGenes vise à conduire une enquête capable de déterminer des macronutriments qui faciliteront la prévention du gain de poids et donc la reprise de poids. L'objectif est d'étudier les interactions entre les composants diététiques et les facteurs génétiques avec les facteurs comportementaux. Avec l'accès à de grandes cohortes prospectives à long terme à travers l'Europe et la présence de données cliniques et nutritionnelles, DiOGenes est une occasion unique d'identifier les interactions gènes-nutriments associées à des changements en termes de masse corporelle et de tour de taille. Pour examiner l'impact des changements dans la composition en macronutriments sur la variation du poids, une étude diététique à long terme de familles entières comprenant à la fois des membres obèses et de poids normal dans 8 pays différents à travers l'Europe a eu lieu.

Une analyse longitudinale à grande échelle de la variation génétique dans les gènes candidats, ainsi que de nouvelles approches telles que l'expression des gènes dans les tissus adipeux et le plasma (péptidomique) donne l'occasion d'identifier des ensembles de polymorphismes de l'ADN, l'ARN messager du tissu adipeux et les peptides du plasma permettant la prédiction de la réponse d'un individu à des nutriments en termes de changement de poids, qui à son tour, permet de proposer un traitement basé sur un régime adéquat. Des études épidémiologiques sont soutenues par une analyse détaillée de l'état psychologique et les réponses comportementales des sujets. Le consortium va se focaliser sur l'identification des principaux prédicteurs du comportement psychologique de la prise de poids pour le diagnostic des risques de prise de poids et pour une meilleure adéquation de

l'alimentation aux besoins des consommateurs. Les technologies alimentaires jouent également un rôle important. Le présent projet comprend des études alimentaires afin de développer des aliments appréciés par les consommateurs mais en même temps capables de renforcer les signaux de satiété. Cela nécessite une combinaison de compétences et de disciplines, et une collaboration entre l'industrie et la science.

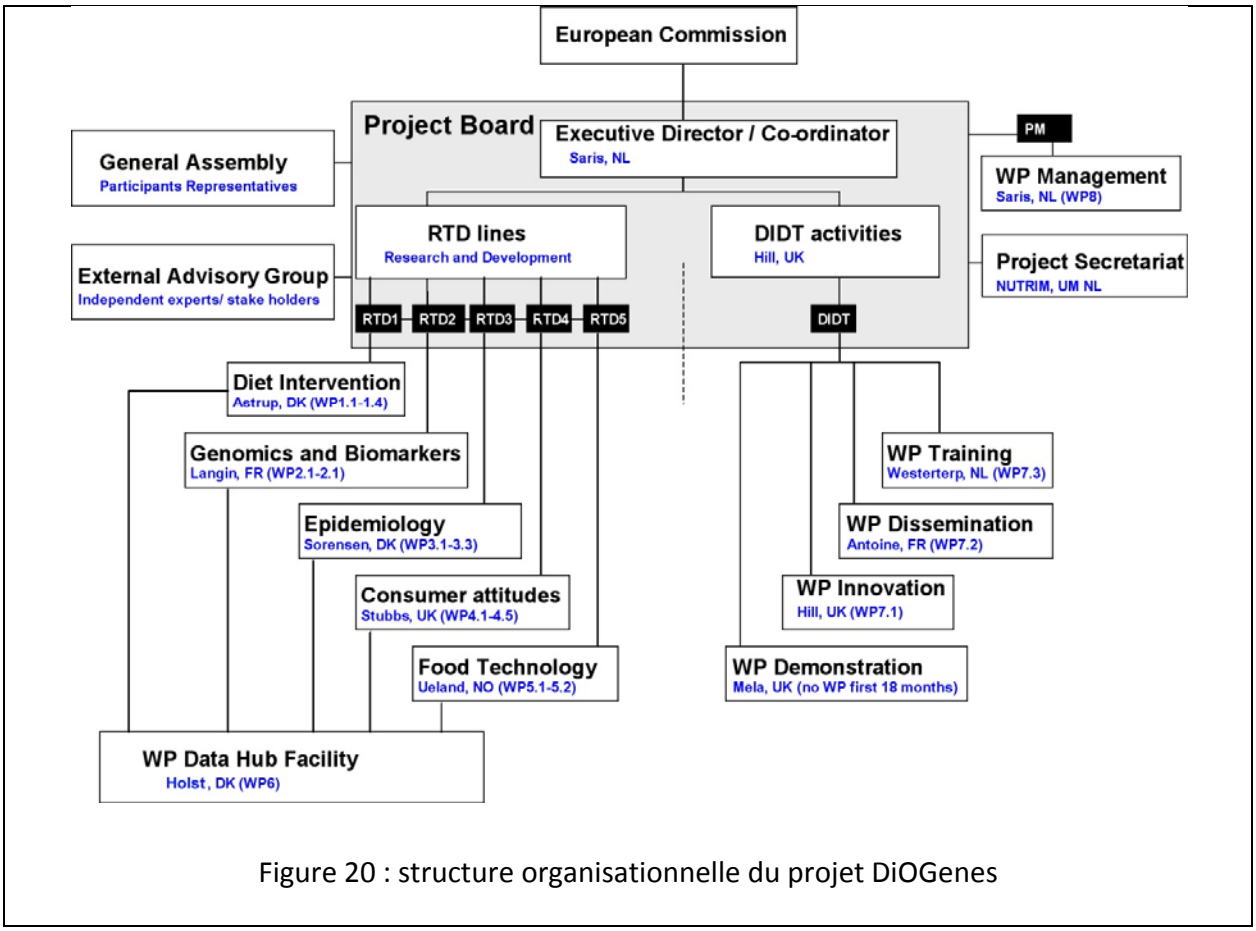


Figure 20 : structure organisationnelle du projet DiOGenes

Dans l'ensemble, l'objectif général du programme intégré du projet DiOGenes est de réduire l'ampleur des problèmes de santé de la surcharge pondérale, de l'obésité et les co-morbidités chez les consommateurs européens. Afin de répondre à cet objectif général, et en conformité avec les lignes directives du programme Européen, DiOGenes se focalisera sur des facteurs génomiques et diététiques et leur rôle dans la prise de poids, la reprise de poids et les co-morbidités associés. En particulier, le consortium vise à identifier, grâce à une approche scientifique intégrative, les facteurs psychologique, des facteurs génétiques ainsi que les

marqueurs biologiques capables de fournir une base scientifique pour prédire si un sujet est capable de maintenir son poids corporel. L'objectif du projet est de fournir et d'accroître le bien-être des consommateurs européens en exploitant les nouvelles connaissances créées pour promouvoir une hygiène alimentaire saine, de haute qualité en minimisant les risques de surpoids et d'obésité.

4.2.2 **Sélection des sujets pour l'analyse prédictive**

Dans le cadre de ce projet européen, 596 sujets ont suivi un régime hypocalorique de 8 semaines et ayant perdu 8% de leurs poids à la fin de ce régime ont été sélectionnés pour cette analyse. Parmi ces sujets, seulement 513 patients avaient les données nécessaires disponibles avant (temps : CID1) et après (temps : CID2) le régime. Un spécimen de graisse sous-cutanée abdominale (~1g) a été obtenu par aspiration sous anesthésie locale. Les biopsies ont été stockées à -80 ° C jusqu'au moment de l'analyse. Pour cette analyse, nous avions besoin d'avoir des données de patients après 48h du début du régime (temps : CID1b) et cela afin d'avoir des modèles de prédiction basés sur une courte variation temporelle. Seulement 79 patients répondent à ce critère et parmi eux 67 ont une quantité d'ARN extraite suffisante pour l'extraction d'ARN. Ces patients proviennent de 3 centres de recrutement (Angleterre, Danemark et Hollande). La présente étude ciblant les femmes, nous avons donc retenu 44 sujets parmi les 79 qui étaient des femmes. Pour construire les groupes, nous avons sélectionné, comme pour Nugenob, la perte de poids (Kg) comme critère, mais nous avons remarqué une différence du BMI au temps CID1 ce qui nous a conduit à choisir la variation en % de la perte de poids comme critère de sélection des groupes. Deux groupes ont été formés pour la variation de la perte de poids après 8 semaines: les «bons répondeurs» (sujets ayant perdu entre 13% et 17% de leur poids initial) et les «mauvais répondeurs» (sujets ayant perdu entre 8% et 10% de leur poids initial). 27 sujets répondent à ces critères au final, parmi eux 10 « bons répondeurs » et 17 « mauvais répondeurs ».Pour les sujets sélectionnés, l'ARN total obtenu à partir de biopsies prises à CID1 et CID1b ont été utilisés dans la présente étude prédictive. La Figure 21, décrit l'évolution de la perte de poids pour les deux groupes de l'étude tout au long du régime alimentaire.

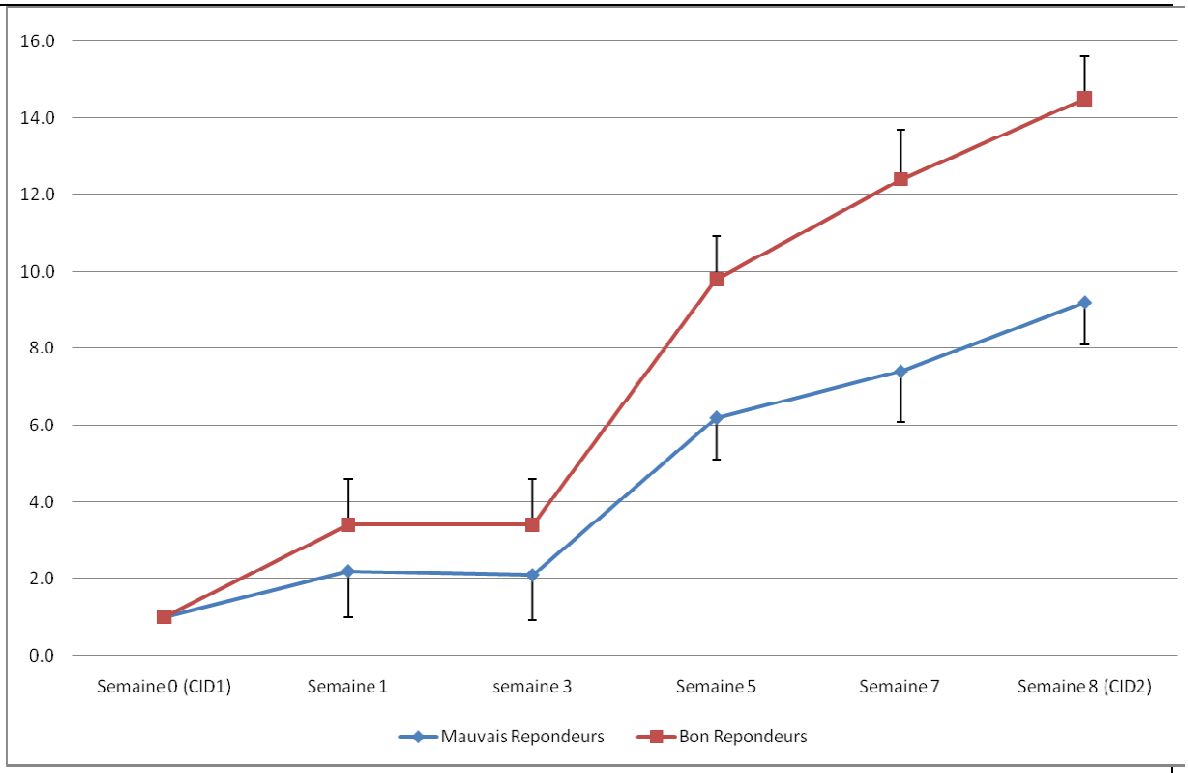


Figure 21: évolution de la perte de poids des sujets de l'étude DiOGenes

4.2.3 **Analyse prédictive à partir des données biopuces**

4.2.3.1 **Prétraitement et préparation des données**

Pour cette étude, nous disposons des données d'expression à CID1 et à CID1b qui proviennent de puces à ADN Agilent de type 4x44K.

À partir de ces données brutes, nous avons procédé à plusieurs prétraitements dans le but de nettoyer les données et les rendre plus exploitables. En un premier temps, nous avons normalisé les données en utilisant le package goulphar (Lemoine, Combes et al. 2006)développé sous R en utilisant la méthode loess, ensuite nous avons enlevé les sondes de contrôle ainsi que les sondes n'ayant pas d'identification. Par la suite, nous avons retiré les gènes ayant des valeurs manquantes en utilisant le package Dprep développé sous R.

A la fin de cette étape nous avons respectivement 18473 et 17202 gènes sans valeurs manquantes à CID1 et CID1b dont 16270 gènes en commun entre CID1 et CID1b.

4.2.3.2 **Prédiction de la perte de poids suite au régime de 8 semaines**

Le but de cette analyse est de prédire la perte de poids après 8 semaines de régime à partir des données d'expressions avant le régime. Ensuite, de déterminer si l'évolution de l'expression à 48 heures après le régime est capable de mieux prédire l'efficacité du régime que les données basales. Pour cela, nous allons construire trois modèles prédictifs de la perte de poids : avec les données CID1, avec les données à CID1b et avec la variation CID1b-CID1.

Pour cette analyse prédictive, nous avons utilisé les méthodes précédemment déployées dans le cadre de l'analyse Nugenob, c'est à dire les K plus proches voisins (Deegalla and Boström 2007), les méthodes d'analyse discriminante (Dudoit, Fridlyand et al. 2002; Diaz-Uriarte and Alvarez de Andres 2006), les forêts aléatoires (Liaw and Wiener 2002; Diaz-Uriarte and Alvarez de Andres 2006) et les machines à vecteurs de supports (Furey, Cristianini et al. 2000; Dudoit, Fridlyand et al. 2002; Diaz-Uriarte and Alvarez de Andres 2006).Tous les algorithmes sont développés sous R: KNN est implémenté dans le package ensemble, DLDA dans le package sma, RF est dans le package randomForest et SVM dans le package Kernlab. Les paramètres standards ont été utilisés pour toutes les méthodes sauf pour SVM, où les paramètres ont été optimisés en utilisant le package svmpath (Trevor Hastie 2004).

Afin d'avoir une estimation robuste, nous avons lancé la validation croisée 10 fois et pris la moyenne des résultats obtenus. La précision de la prédiction des modèles a été comparée à l'approche naïve qui place tous les sujets dans la classe majoritaire (pour les données de l'obésité, la précision de la prédiction de l'approche naïve est de 63%), un résultat significativement supérieur à celui de l'approche naïve a été considéré comme une amélioration de la précision de prédiction.

En utilisant l'intégralité de l'expression des 16270 gènes des données recueillies à CID1, nous obtenons une amélioration de la précision de la prédiction par rapport à l'approche de référence comme l'indique la Figure 22. Le meilleur résultat est obtenu par la méthode DLDA

avec une précision de la prédiction 95% ± 0%. KNN, RF et SVM (noyau linéaire/radial) prédisent la perte de poids avec une précision de 90% ± 0%, 89,5% ± 1,6%, 91,3% ± 3,3% et 86,5% ± 4,6% respectivement.

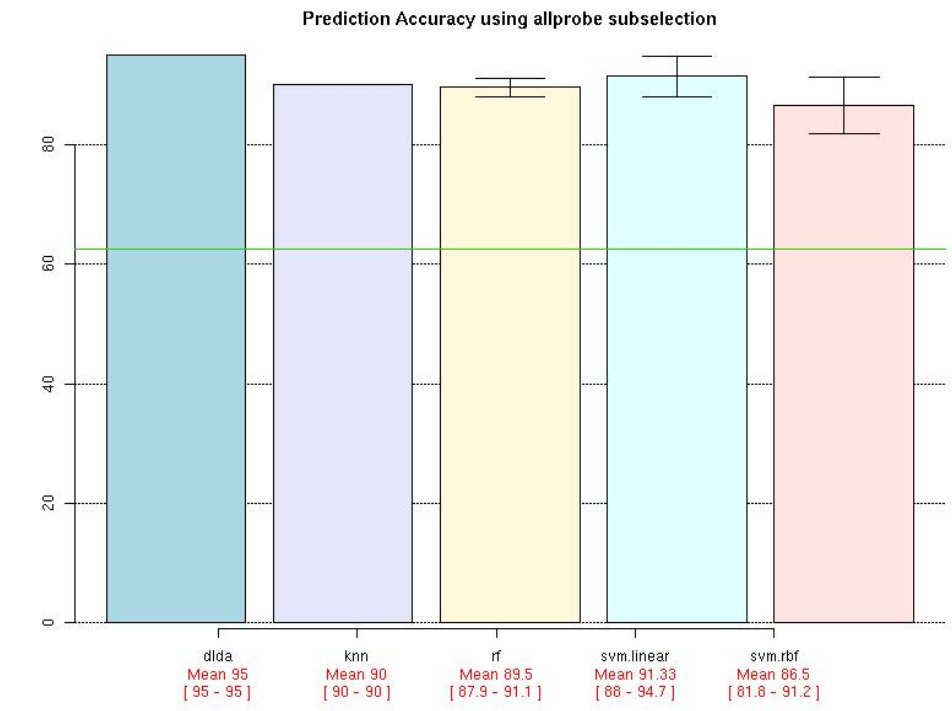


Figure 22: Résultat de la prédiction de la perte de poids à partir des données CID1, la ligne horizontale en vert est celle du classeur naïf.

Avec l'intégralité de l'expression des 16270 gènes des données recueillies à CID1b, les résultats restent supérieurs à ceux de l'approche naïve comme l'indique la Figure 23. Le meilleur résultat est obtenu par la méthode SVM (noyau linéaire) avec une précision de la prédiction 90,2% ± 4,7%. DLDA, KNN, RF et SVM (noyau radial) prédisent la perte de poids avec une précision de 83.5% ± 8,5%, 89,2% ± 2,3%, 87,8% ± 9,8% et 72% ± 9,4% respectivement.

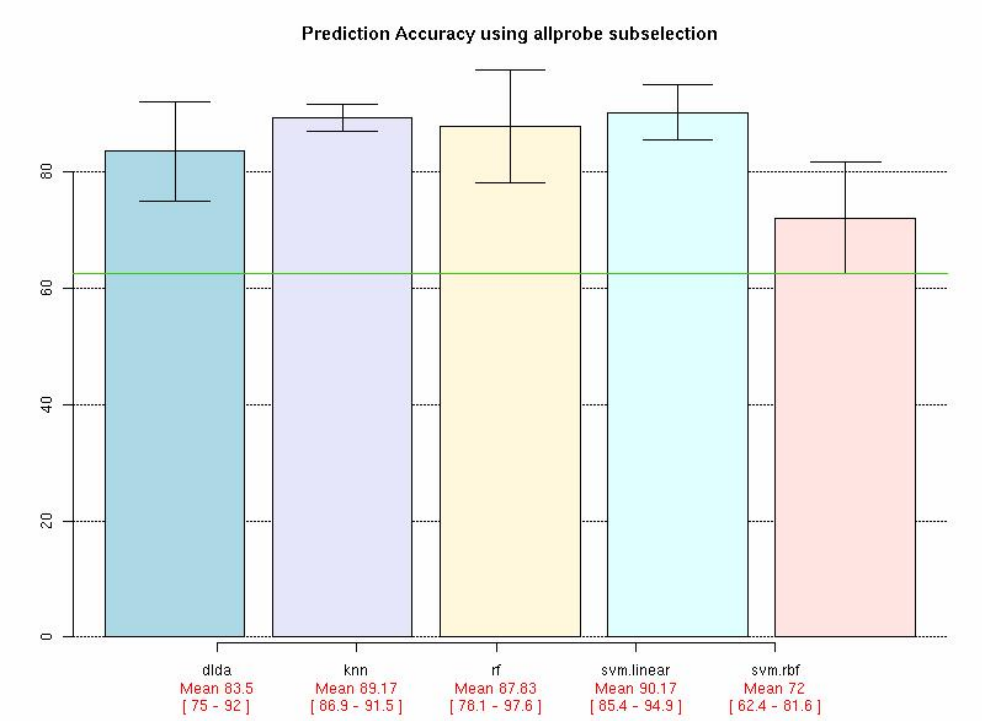


Figure 23: Résultat de la prédiction de la perte de poids à partir des données CID1b, , la ligne horizontale en vert est celle du classeur naïf.

En prenant l'intégralité de la variation des gènes entre CID& et CID1b de l'expression des 16270 gènes, les résultats restent supérieurs a ceux de l'approche naïve dans ce cas aussi comme l'indique la Figure 24. Le meilleur résultat est obtenu par la méthode SVM (noyau linéaire) avec une précision de la prédiction 94% ± 1,7%. DLDA, KNN, RF et SVM (noyau radial) prédisent la perte de poids avec une précision de 88% ± 6,3%, 84,8% ± 3,9%, 90% ± 4,1% et 76,7% ± 3,7% respectivement.

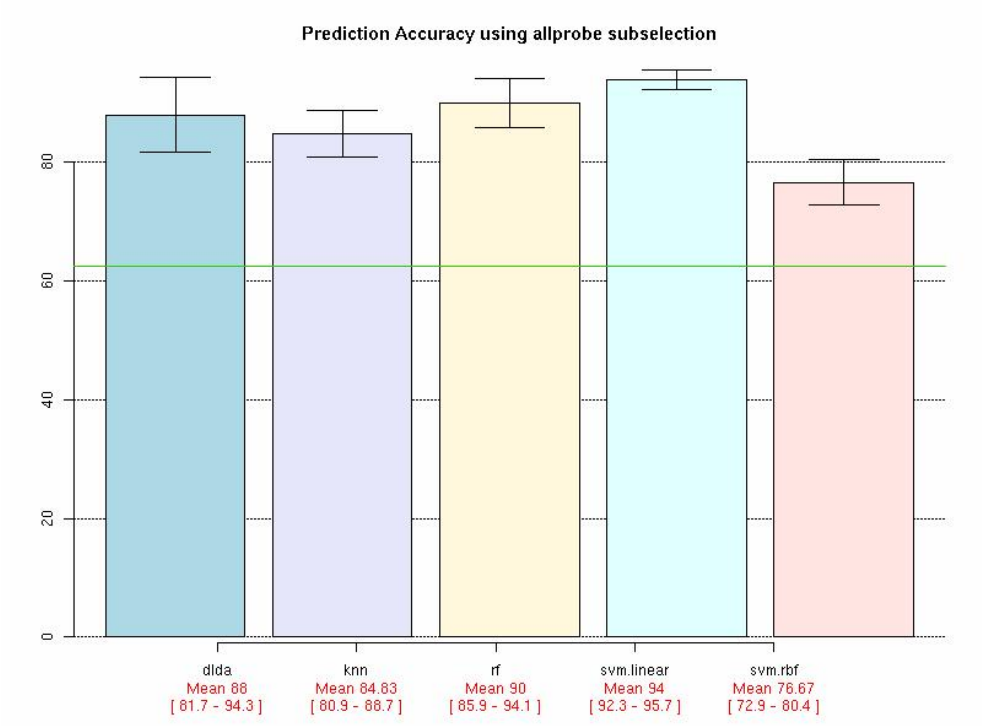


Figure 24: Résultat de la prédiction de la perte de poids à partir de la variation CID1b-CID1, la ligne horizontale en vert est celle du classeur naïf.

La Table 15 résume les résultats obtenus à partir des différentes données utilisées, les meilleurs résultats sont obtenus avec les données à CID1 en appliquant la méthode DLDA. Nous obtenons des résultats similaires avec les données de variation et la méthode SVM avec un noyau linéaire.

Table 15 : Résultats de la précision de la prédiction avec les données de l'obésité

Données utilisées	DLDA	KNN	RF	SVM.L	SVM.R
CID1	95%±0%	90%±0%	89,5%±1,6%	91,3%±3,3%	86,5%±4,6%
CID1b	83.5%±8,5%	89,2%±2,3%	87,8%±9,8%	90,2% ± 4,7%	72%±9,4%
CID1b-CID1	88% ± 6,3%	84,8%±3,9%	90%±4,1%	94% ± 1,7%.	76,7%±3,7%

Nous avons aussi essayé de déterminer l'effet de la sélection de variable sur les résultats de la prédiction. Nous avons pris les 10 meilleurs gènes en validation croisée par la méthode de Fisher ou un test-t, ces résultats sont indiqués dans laTable 16.

Table 16 : résultat de la précision de la prédiction avec les données de l'obésité en appliquant la sélection de variables

Données utilisées	Sélection de gènes	DLDA	KNN	RF	SVM.L	SVM.R
CID1	Top10 par Fisher-score	84,7%±6,1%	85,3%±6,0%	83,2%±5,2%	86,7%±6,7%	80,2%±8,2%
	Top10 par T-Test	87%±5,1%	84,8%±7,2%	81,8%±4,8%	85,7%±7,5%	84%±6,6%
CID1b	Top10 par Fisher-score	76,2%±6,5%	69,7%±10,6%	71,8%±7,4%	70,5%±6,9%	68,3%±9,0%
	Top10 par T-Test	88,7%±7,7%	85,8%±9,0%	82,2%±11,6%	87,5%±6,9%	87,3%±5,3%
CID1b-CID1	Top10 par Fisher-score	82,2%±6,9%	76,3%±10,8%	69,8%±11,1%	80%±7,9%	76,2%±7,8%
	Top10 par T-Test	92,5%±7,5%	89,5%±1,6%	90,2%±8,9%	90,3%±9,1%	90,3%±7,9%

Les résultats obtenus par ces deux méthodes de sélection n'ont pas abouti à une amélioration de la prédiction comme l'indique la Table 16, cependant avec les données de variation CID1b-CID1 l'écart de performances en moyenne est réduit. Ceci est intéressant puisque nous avons des résultats proches avec un nombre réduit de 10 gènes au lieu de 16270. Cette sélection réduite est plus pratique d'un point de vue clinique et biologique car plus facile à étudier.

Nous avons par la suite essayé de déterminer les gènes qui varient entre CID1 et CID1b et considérer cette sous liste pour la prédiction, en partant de l'hypothèse que les gènes qui varient entre CID1 et CID1b seraient ceux les plus impliqués dans le mécanisme génétique de perte de poids. Pour cela nous avons considéré CID1 et CID1b comme étant deux conditions et nous voudrions connaitre les gènes qui prédiraient le changement entre ces deux classes. Pour ce faire, nous avons utilisé le package **`pamr`** développé sous R qui permet de déterminer des listes prédictives classées par score. Nous reportons dans la table 12, les seuils pour lesquels nous avons un changement de la précision.

Table 17 : nombre de probe et précisons pour chacun des seuils déterminés par pamR

Seuil	Nombre de probe	Précision
0	16270	83,33%
0,820508656	6174	83,33%
1,025635819	4422	85,42%
1,435890147	2104	85,42%
1,641017311	**1480**	87,50%
5,538433425	**12**	87,50%
5,743560589	2	77,08%

Les meilleurs résultats, exprimant une bonne discrimination entre CID1 et CID1b, sont obtenus en sélectionnant un seuil entre 1,64 et 5,53 c.à.d. en prenant les meilleures listes de gènes regroupant entre les 1480 et 12 gènes. Nous avons testé 2 sélections en prenant des listes de 12 et 3 gènes. Les résultats sont donnés dans la Table 18.

Nous remarquons que l'expression des meilleurs 12 gènes par pamr de CID1 permet d'avoir des résultats très proches de ceux obtenus à partir de la totalité des gènes avec des petites variabilités en comparaison avec les listes du test-t et de fisher.

Table 18 : résultat de la précision de la prédiction avec les sélections pamr CID1 versus CID1b

Données Utilisées	Sélection de gènes	DLDA	KNN	RF	SVM.L	SVM.R
CID1	Top3 par pamr	86,8%±4,7%	86%±3,2%	86%±3,2%	86%±3,2%	86,8%±3,9%
	Top12 par pamr	92%±4,3%	91,5%±3,9%	90,5%±3,1%	91%±3,2%	88%±5,7%
CID1b	Top3 par pamr	27,5%±13,7%	61,3%±6,9%	64,3%±4,4%	47,7%±2,3%	57,7%±19,8%
	Top12 par pamr	34,3%±10,8%	65%±6,7%	58,5%±12,3%	64,3%±6,6%	58,8%±9,9%
CID1-CID1b	Top3 par pamr	91,2%±5,3%	80,7%±6 ,2%	80%±8,4%	83,7%±3%	81%±5,6%
	Top12 par pamr	90%±4,6%	86,7%±3%	86,8%±3,3%	86,5%±2,9%	82,5%±4,4%

D'après les résultats présentés ci-dessus, nous pouvons dire que le temps CID1 est suffisant pour répondre aux problèmes de prédiction et que les données à 48h ainsi que la variation entre les deux temps ne fournissent par d'amélioration significative de la prédiction. Dans ce qui suit, nous allons nous intéresser uniquement aux modèles construits à partir de CID1.

4.2.3.3 **Effets de la randomisation sur les résultats de la prédiction**

Dans ce qui suit, nous avons effectué des permutations sur les labels afin de voir si une prédiction d'un problème aléatoire serait d'une même qualité ou bien que celle-ci dégraderait les performances de prédiction. Cela a été effectué sur les données complètes (16270 gènes) ainsi que la sélection pamr (12 gènes).

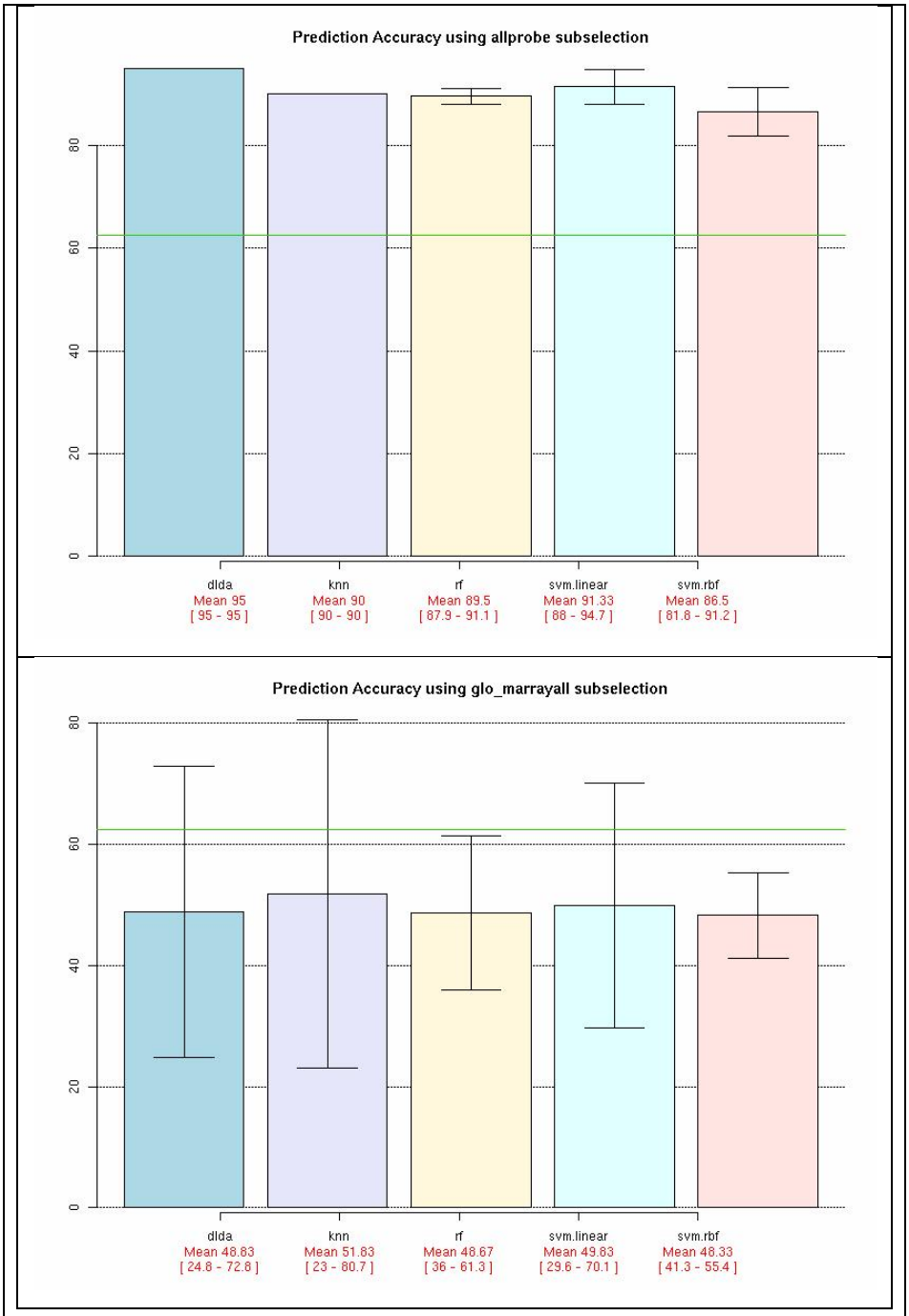


Figure 25 : Effet de la randomisation sur la prédiction à partir de l'expression des 16270 gènes à CID1

Nous constatons que la 'randomisation' détériore les performances de la prédiction et cela indépendamment de l'algorithme choisi, que ce soit en prenant l'intégralité des gènes (Figure 25) ou bien la sélection pamr (Figure 26). Les résultats sont moins bons que ceux

obtenus par l'algorithme naïf. Ceci permet d'induire que les données contiennent une information capable de différentier et prédire la perte de poids.

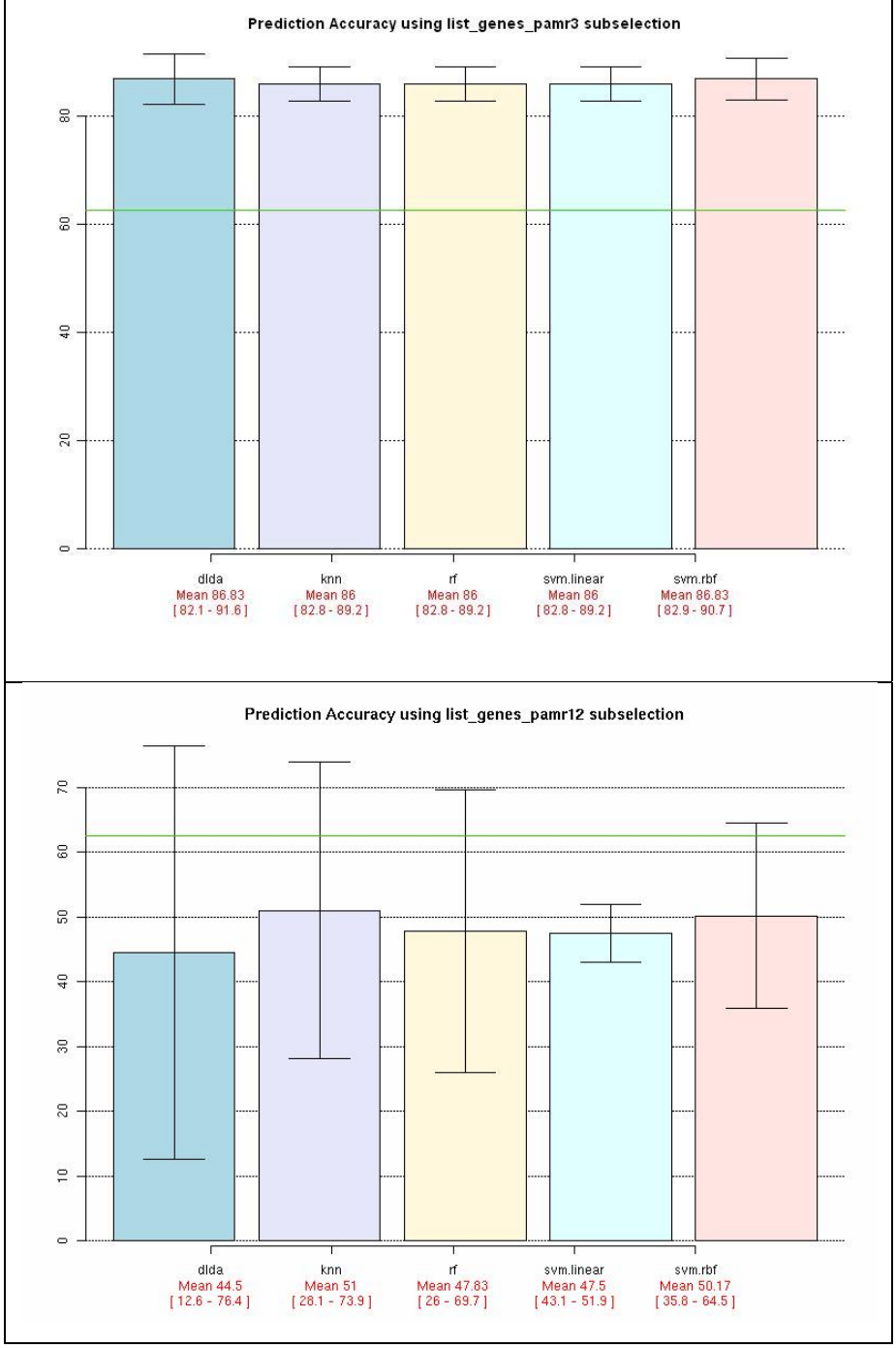


Figure 26 : effet de la randomisation sur la prédiction à partir de l'expression de la sélection des 12 meilleurs gènes par pamr à CID1

4.2.4 **Discussion**

La prédiction de la perte de poids avant un régime pourrait permettre de mieux aider les praticiens dans la définition de la prise en charge du patient. Les résultats de la présente analyse sont prometteurs et montrent que l'utilisation de la transcriptomique pourrait contribuer activement à la mise en place d'un système d'aide à la décision thérapeutique pour la prédiction de la perte de poids suite à un régime. La sélection obtenue pourrait rendre le système plus simple, puisqu'il suffit de regarder une sélection d'une dizaine de gènes pour avoir une estimation du taux de réussite du régime préconisé.

rank	PrimaryAccession	EntrezGeneID	GeneSymbol	GeneName	Cytoband
1	NM_032805	84891	ZSCAN10	zinc finger and SCAN domain containing 10	hs|16p13.3
2	THC2723346				hs|1p34.3
3	A_24_P714316				hs|22q11.21
4	XM_929673	54553	DKFZP434I0714	hypothetical protein DKFZP434I0714	hs|4q31.3
5	NM_001080495	84629	KIAA1856	KIAA1856 protein	hs|7p22.1
6	NM_022454	64321	SOX17	SRY (sex determining region Y)-box 17	hs|8q11.23
7	A_24_P42107				hs|11p15.5
8	NM_015366	55615	PRR5	proline rich 5 (renal)	hs|22q13.31
9	BM547196				hs|2q11.2
10	THC2538856				hs|5q35.3
11	NM_178438	254910	LCE5A	late cornified envelope 5A	hs|1q21.3
12	AK092271				hs|9q33.3

Table 19: liste des 12 meilleurs gènes par pamR

En regardant le fichier d'annotation des gènes de la sélection faite par Pamr, on retrouve uniquement 6 gènes annotés parmi les 12. Les 6 gènes reconnus sont les suivants :

- zinc finger and SCAN domain containing 10 [ZSCAN10]
- hypothetical protein DKFZP434I0714
- KIAA1856 protein [TNRC18]
- SRY (sex determining region Y)-box 17 [SOX17]
- proline rich 5 (renal) [Prr5]
- late cornified envelope 5A[LCE5A]

Pour avoir plus d'information sur la liste des 12 gènes, nous avons essayé d'exploiter les bases de données biologiques afin d'extraire plus d'information sur la liste de gènes. Pour cela, nous avons utilisé l'outil Blast du NCBI qui permet, à partir de la position chromosomique d'un gène, de connaitre les informations sur ce dernier. L'avantage est que cette base est constamment mise à jour et par conséquent nous pourrions obtenir des informations sur les gènes pour lesquels des informations étaient absentes dans le fichier d'annotation du constructeur des puces à ADN. Comme nous pouvons le constater dans la Table 20, nous avons pu obtenir des informations sur tous les gènes de notre liste à une exception près. Nous remarquons aussi que 2 gènes différents peuvent coexister dans une même position chromosomique, d'où une liste de 17 gènes au lieu de 12 initialement.

rank	PrimaryAccession	EntrezGene ID	gene_name	gene name
1	NM_032805	84891	ZSCAN10	Zinc finger and SCAN domain containing 10
2	THC2723346	6421	SFPQ	splicing factor proline/glutamine-rich
3	THC2723346	9202	ZMYM4	zinc finger, MYM-type 4
4	A_24_P714316	5413	SEPT5	Septin 5
5	A_24_P714316	2812	GP1BB	glycoprotein Ib (platelet), beta polypeptide
6	XM_929673	54553	DKFZP434I0714	Hypothetical protein DKFZP434I0714
7	NM_001080495	84629	KIAA1856	TNRC18:Trinucleotide repeat containing 18

8	NM_022454	64321	SOX17	SRY (sex determining region Y)-box 17
9	A_24_P42107	NA		
10	NM_015366	55615	PRR5	Rho GTPase activating protein 8
11	BM547196	80705	TSGA10	Testis specific, 10
12	BM547196	343990	C2orf55	chromosome 2 open reading frame 55
13	THC2538856	54540	FLJ10404	Hypothetical protein FLJ10404
14	THC2538856	54732	TMED9	Transmembrane emp24 protein transport domain containing 9
15	NM_178438	254910	LCE5A	Late cornified envelope 5A
16	AK092271	4010	LMX1B	LIM homeobox transcription factor 1, beta
17	AK092271	23099	ZBTB43	Zinc finger and BTB domain containing 43

Table 20:Liste étendue obtenue à partir de blast des la liste des probe id obtenue par pamr

L'étape suivante consiste à déterminer, à partir de la littérature, la fonction ainsi que des informations relatives à ces gènes dans des analyses déjà publiées. Nous avons utilisé le portail SOURCE de l'université de standford afin de récolter ces informations.

rank	gene_name	function	littérature
1	ZSCAN10	Transcription factor activity	
2	SFPQ	Transcription / Regulation	late splicing factor, repressing the transcription of multiple oncogenic genes/acting as a progesterone receptor corepressor and contributing to the functional withdrawal of progesterone and the initiation of labor/repressor which interacts with SIN3A and mediates silencing through the recruitment of HDACs to the receptor DNA binding domains

3	ZMYM4	Zinc ion/Metal / DNA Binding	he 3'-UTR region of the mRNA encoding this protein contains a motif called CDIR (for cell death inhibiting RNA) that binds HNRPD/AUF1 and HSPB1/HSP27. It is able to inhibit interferon-gamma induced apoptosis
4	SEPT5	involved in myeloid leukemogenesis / regulating cytoskeletal organization and cytokinesis /may be playing a functional role in platelet granular secretion	PNUTL1 and GP1BB are encoded on the same DNA stand, overlapping with the same transcriptional orientation
5	GP1BB		
6	DKFZP434I0714		
7	KIAA1856	DNA binding	
8	SOX17	Transcription factor binding,	
9			
10	PRR5	regulates platelet-derived growth factor receptor beta expression and signaling	interacts with RICTOR, but not RAPTOR, and the interaction is independent of FRAP1 and not disturbed under conditions that disrupt the FRAP1-RICTOR interaction (Rictor-binding subunit of mTORC2)
11	TSGA10		tumour-associated antigen of cutaneous lymphoma / testis-expressed protein with a key role in spermatogenesis
12	C2orf55		
13	FLJ10404		
14	TMED9	involved in vesicular protein trafficking	
15	LCE5A		

16	LMX1B	Transcription factor activity / Binding	involved in dorsal-ventral limb patterning and kidney development and patterning modulator of motor axon guidance / essential for the specification of dorsal limb fates at the zeugopodal and autopodal level in vertebrates / regulation of the coordinated expression of COL4A3 and COL4A4 required for normal glomerular basal membrane morphogenesis / required for the differentiation and migration of neurons within the dorsal spinal cord
17	ZBTB43	binding	involved in the regulation of transcription, DNA-dependent

Table 21: fonction et littératures de la sélection étendue des gènes cibles

Nous avons ensuite analysé le comportement de ces gènes dans les différentes expérimentations conduites dans l'équipe 7 de l'UMRS 872 :

- Caractérisation des profils transcriptionnels des principales fractions cellulaires du tissu adipeux humain (Adipo VS. SVF) : Le but de cette l'analyse était d'une part de caractériser le profil fonctionnel individualisant chacune des deux fractions, et d'autre part de décrire les relations reliant les processus cellulaires mobilisés spécifiquement dans chaque fraction. Des prélèvements de tissu adipeux sous-cutanés réalisés chez 9 sujets de sexe féminin (BMI 27,9 ± 6,8 kg/m²) ont été soumis à une digestion enzymatique, suivi par une ultracentrifugation, afin de séparer les fractions cellulaires du tissu (Cancello, Henegar et al. 2005). Après extraction de l'ARN, 6 puces à ADNc (Stanford) ont été réalisées afin de mesurer l'expression transcriptionnelle différentielle dans les deux fractions. L'analyse de l'expression différentielle, réalisée avec l'outil SAM en fixant un seuil de FDR ≤ 5%, a permis d'identifier 3492 transcrits exprimés

principalement dans l'adipocyte mature et 4215 transcrits exprimés de façon prédominante dans la SVF.

- Caractérisation du profil transcriptionnel du tissu adipeux de sujets obèses en phase pondérale stable (Obtem): Les données analysées dans cette situation proviennent de mesures d'expression transcriptionnelle du tissu adipeux total réalisées au moyen de puces pangénomiques à ADNc (Stanford) chez 25 sujets présentant une obésité massive en poids stable (BMI 40,57 ± 7,9 kg/m²) et 10 témoins normopondéraux (BMI 23,67 ± 1,51 kg/m²) sains. Cette condition expérimentale a eu comme but de caractériser les perturbations affectant la signature transcriptionnelle du tissu adipeux au cours de l'obésité massive. L'analyse de l'expression différentielle des clones ADNc récupérés dans au moins 80% des expériences, réalisée avec l'outil SAM en considérant un seuil de la FDR ≤ 5%, a permis d'identifier 366 transcrits sur- et 474 transcrits sous-exprimés. L'analyse fonctionnelle de ces transcrits a permis d'identifier 704 (307 sur- et 397 sous-exprimés) annotés par des catégories Gene Ontology (GO) et 253 (101 suret 152 sous-exprimés) annotés par des catégories appartenant à Kyoto Encyclopedia of Genes and Genomes (KEGG).
- Caractérisation du profil transcriptionnel du tissu adipeux de sujets obèses après perte de poids induite par chirurgie bariatrique (Bypass) : il a été montré auparavant que la perte pondérale est accompagnée par une amélioration du profil inflammatoire, associée à une diminution de l'infiltration macrophagique qui affecte le WAT de sujets obèses (Clement, Viguerie et al., 2004). Pour mieux caractériser l'association entre la variation de la masse adipeuse, les phénomènes inflammatoires locaux et le remodelage structural du tissu, l'équipe a examiné le profil fonctionnel de la signature transcriptionnelle du WAT de sujets obèses après une perte significative de poids induite par une chirurgie bariatrique (c.-à-d. by-pass gastrique). Des mesures du profil d'expression du tissu adipeux total ont été réalisées chez 10 sujets massivement obèses (BMI 47,65 ± 4,4 kg/m), avant et 3 mois après une perte de poids significative induite par la chirurgie bariatrique. L'analyse de l'expression différentielle des sondes d'ADNc récupérées dans au moins 80% des expériences, réalisée avec l'outil SAM en considérant

un seuil de la FDR ≤ 5%, a permis d'identifier 1744 transcrits sur- et 1627 transcrits sous-exprimés. L'analyse fonctionnelle de ces transcrits a identifié 2687 gènes (1390 sur- et 1297 sous-exprimés) annotés par des catégories GO et 868 gènes (450 sur- et 418 sous-exprimés) annotés par des catégories KEGG.

- Caractérisation du profil transcriptionnel des pré-adipocytes humains cultivés avec des milieux conditionnés de macrophages, différenciés à partir de monocytes circulants, et étudié leur phénotype, ainsi que leurs capacités de synthèse des protéines structurales sous l'effet des facteurs pro-inflammatoires d'origine macrophagique : la quantification du degré de fibrose interstitielle du WAT a été réalisée chez 10 sujets affectés par une forme d'obésité morbide avant et 3 mois après une intervention de chirurgie bariatrique, ainsi que dans 10 témoins normopondéraux appariés pour l'âge et le sexe. Un degré de fibrose interstitielle significativement plus important dans le WAT de sujets obèses que celui observé chez les témoins normopondéraux (6,29%±2 vs. 2,19% ±0,25, p-value ≤0,05 ; Fig. 52D) a été observé. Aucune variation significative du degré de fibrose interstitielle n'a pu être mise en évidence 3 mois après la chirurgie bariatrique, en dépit d'une tendance globale à la diminution. La synthèse de collagène de type I et de la fibronectine par les préadipocytes humains sous l'influence des facteurs pro-inflammatoires existants dans le milieuconditionné de macrophages. Ces résultats suggèrent que le pré-adipocytes humains, en combinaison avec d'autres cellules de la SVF, pourraient contribuer à la synthèse excessive de composants structuraux dans le WAT de sujets obèses, sous l'effet du microenvironnement inflammatoire local.

Nous avons évalué la liste de gènes en notre possession et nous avons obtenu le tableau suivant :

rank	gene_name	adipo_svf	adipo_svf_rank	obtem	obtem_rank	bypass	bypass_rank	mac	mac_rank
1	ZSCAN10								
2	SFPQ	-	**7384**			+	**13**		
3	ZMYM4							-	**43**
4	SEPT5								
5	GP1BB								
6	DKFZP434I0714	-	**329**			-	**2531**		
7	KIAA1856							-	**219**
8	SOX17	-	**5462**						
9	?								
10	PRR5								
11	TSGA10								
12	C2orf55	-	**6085**						
13	FLJ10404								
14	TMED9								
15	LCE5A								
16	LMX1B								
17	ZBTB43								

Table 22: classement des genes dans les experiences internes de l'équipe

A partir de ce tableau, nous constatons que le gène SFPQ (splicing factor proline/glutamine-rich) est surexprimé chez les obeses. Ce gène est classé 13^{ième} dans la liste des gènes différentiellement exprimés. Le gène ZMYM4 (zinc finger, MYM-type 4) est sous exprimé dans l'expérimentation des macrophages et il est classé en position 43.

4.3 Bilan des résultats transcriptomiques : Nugenob Versus Diogenes.

Nous avons présenté dans ce qui précède deux études de prédiction de la perte de poids suite à une restriction calorique. Dans la première analyse « nugenob » la durée du régime est de 10 semaines. Le critère fixé pour le succès du régime est que le patient perde entre 8 et 12 kg de son poids initial contre une perte inférieure à 4 kg pour l'échec du régime. Dans la deuxième analyse qui concerne le projet « Diogenes », nous nous sommes intéressés aux patients ayant réussi un régime de 8 semaines et nous voudrions discriminer un bon répondeur (perte de poids entre 13% et 17%) d'un mauvais répondeur (entre 8% et 10%). Nous avons pu constater à travers ces deux analyses une différence significative entre les résultats obtenus, comme le montre la Figure 27.

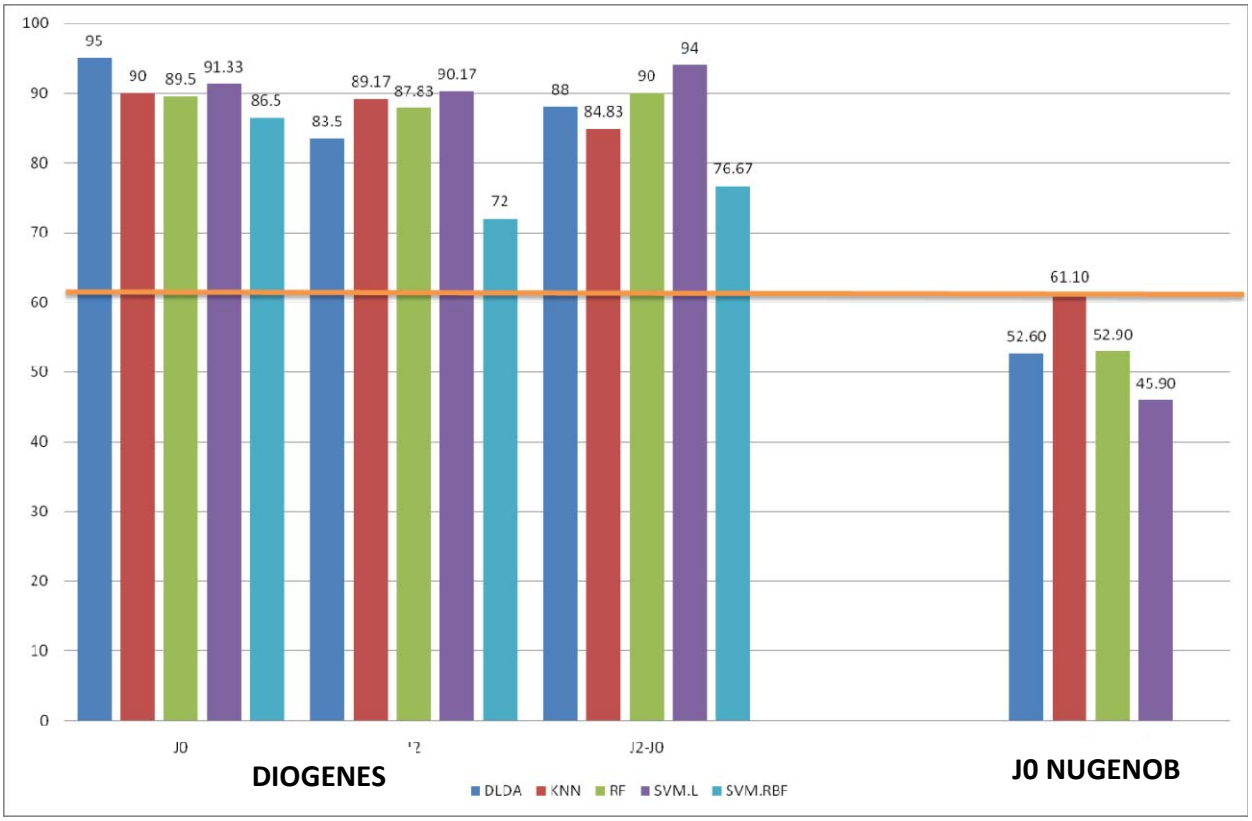


Figure 27: comparaison des résultats de prédiction entre Nugenob et Diogenes

Nous observons dans la première analyse, des prédicteurs ayant des performances limitées. Par contre, la deuxième analyse permet d'avoir des modèles de prédiction assez

performants comparés aux premiers, ce qui est tout de même surprenant. Nous nous attendons à ce que la différence dans l'expression des gènes entre des personnes ayant perdu du poids (8-12kg) et des personnes ayant perdu peu (moins de 4kg) soit plus marquée que dans la deuxième expérimentation où toutes les personnes ont perdu du poids (par rapport au premier critère de classification) mais cette hypothèse est contredite par les résultats empiriques que nous avons retrouvés. Nous avons alors essayé de comprendre les raisons qui pourraient être à l'origine d'une telle différence dans les résultats, donc nous avons listé dans la Table 23les différents facteurs expérimentaux et techniques capables de contribuer à une modification des résultats.

	Nugenob	Diogenes
Type des puces à ADN	Agilent 44K	Agilent 4x44K
Qualité de l'ARN	Bonne qualité	Bonne qualité
soustraction du bruit de fond	Non	Non
Normalisation	Loess	Loess
Prétraitement	Standard	Standard
Gènes différentiellement exprimés (FDRà 5%)	0	3600
Pool de référence	Pool tissu adipeux	FirstChoice® Human Total RNA Survey Panel(AM6000)
Nombre de centres cliniques	12	3
Choix de la classe	(-) perte < 4kg / (+) perte entre 8 et 12 Kg	(-) perte entre 8% et 10% / (+) perte entre 13%et 17%
Durée du régime	10 semaines	8 semaines

Table 23 : Comparatif des conditions expérimentales entre l'analyse Nugenob et Diogenes

Pour les deux analyses, nous avons utilisé les puces à ADN pangénomique de type Agilent. Dans Nugenob des puces 44K et pour diogenes 4x44K, les dernières sont une optimisation des premières où l'on peut spotter 4 expérimentations sur une même lame, mais c'est toujours la même technologie que les 44k. La qualité d'ARN était bonne dans les deux analyses. La normalisation et le traitement du bruit de fond et le prétraitement sont appliqués de la même manière aux deux analyses. Nous remarquons lors de l'étude des gènes différentiellement exprimés que pour une FDR à 5% aucun gène n'apparait dans ces conditions

pour l'analyse Nugenob alors que pour Diogenes la liste compte 3600 gènes, ce qui veut dire la séparation dans Nugenob entre les deux groupes n'est pas claire, alors que dans Diogenes il existe une différence d'expression visible entre les deux groupes qui ont été définis.

Le pool de référence utilisé dans Diogenes est un pool de référence commercial constitué d'un mélange de tissus (tissu adipeux, cœur, foie, muscle squelettique, l'intestin grêle) alors que celui utilisé dans Nugenob est un pool de tissu adipeux utilisé au sein du laboratoire. Est-ce que ce changement de pool est à l'origine de la modification de la performance entre les deux analyses ? L'idéal pour tester cette hypothèse serait de refaire les puces Nugenob en prenant comme pool de référence le pool commercial. Mais à cause du coût élevé des puces à ADN cette solution n'est pas faisable. Nous avons alors réalisé une lame dite 'jaune' qui mesure le rapport entre les deux pools (pool Diogenes versus pool Nugenob). Cette lame, moyennant des transformations algébriques, permettra virtuellement de générer des données Diogenes avec le pool du tissu adipeux utilisé dans Nugenob. Les données Diogènes normalisées sont sous la forme $DG_{nz} = \log_2\left(\frac{DonnéeDiogenes}{poolDiogenes}\right)$. En un premier temps, nous normalisons la lame jaune et nous obtenons. $LJ_{nz} = \log_2\left(\frac{PoolDiogenes}{PoolNugenob}\right)$. Ensuite, nous transformons les données diogenes normalisées et la lame jaune, nous multiplions les ratios puis nous appliquons de nouveau une transformation logarithmique nous obtenons ainsi les données Diogenes avec le pool de Nugenob $Diogenes_{poolNugenob} = \log_2\left(2^{DG_{nz}} * 2^{LJ_{nz}}\right)$. Une fois ces transformations effectués, nous avons relancé notre analyse avec ces données et comparé les résultats obtenus à partir des données Diogenes avec un pool Nugenob avec ceux obtenus précédemment avec le pool commercial.

La Figure 28 montre que les résultats sont presque identiques et qu'il n'existe pas de différence significative entre les résultats obtenus à partir du pool Diogènes et ceux obtenus à partir du pool Nugenob, ce qui nous amène à la conclusion que le changement de pool n'est pas la cause de cette amélioration des résultats dans l'étude Diogenes.

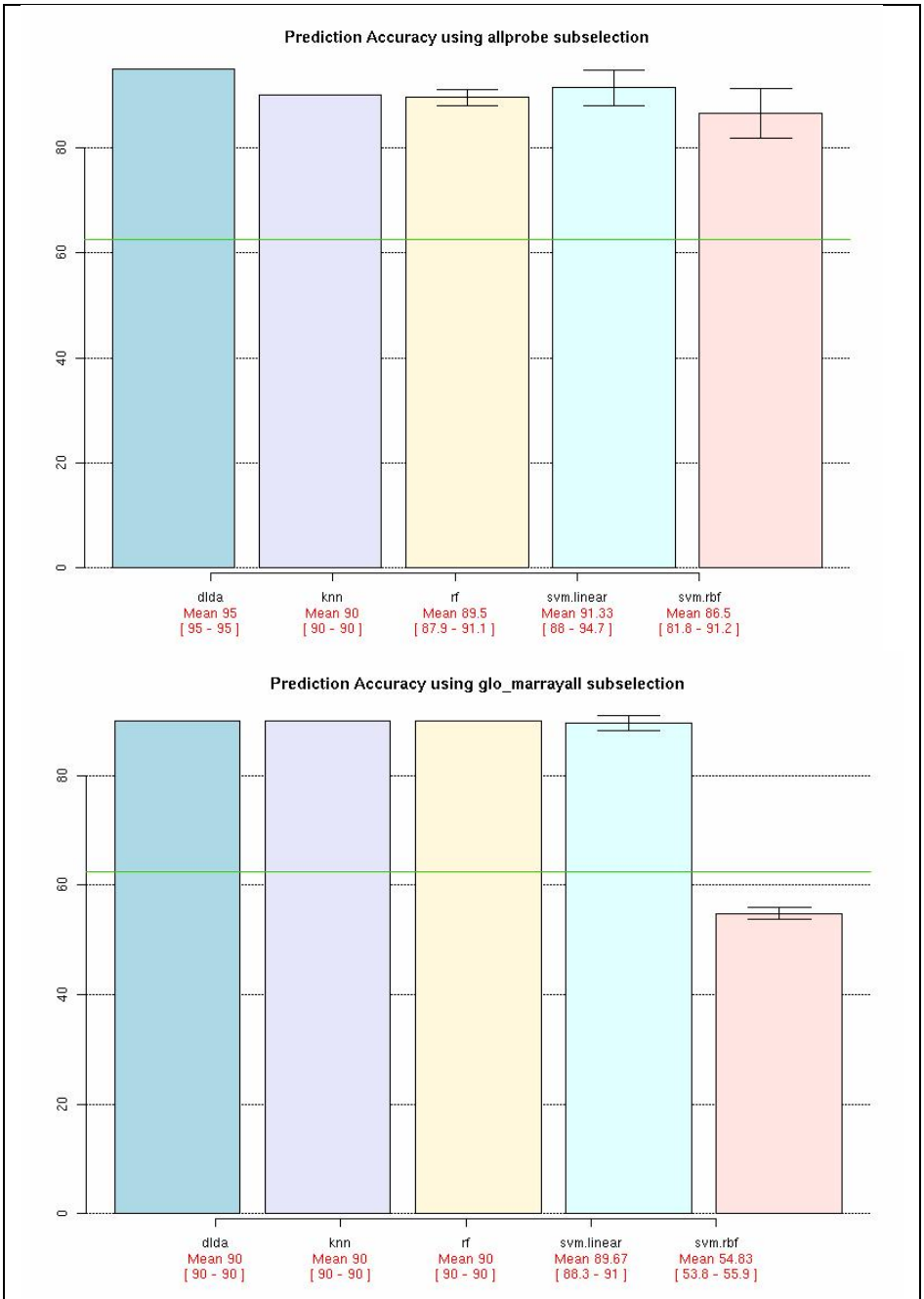


Figure 28 : Comparatif des resultats de prediction à partir des données diogenes avec un pool commercial et avec un pool interne

Dans la Figure 29, nous avons transformé les pertes de poids en pourcentage pour la première analyse et nous reportons la distribution de la variation du poids dans les deux études. La séparation entre les groupes dans Nugenob était visible lorsque le critère était la perte de

poids en Kg, en effectuant la transformation en perte en % nous observons que les deux groupes sont proches. En revanche, dans Diogenes il ya tout de même une bonne séparation entres les mauvais répondeurs et les bons répondeurs.

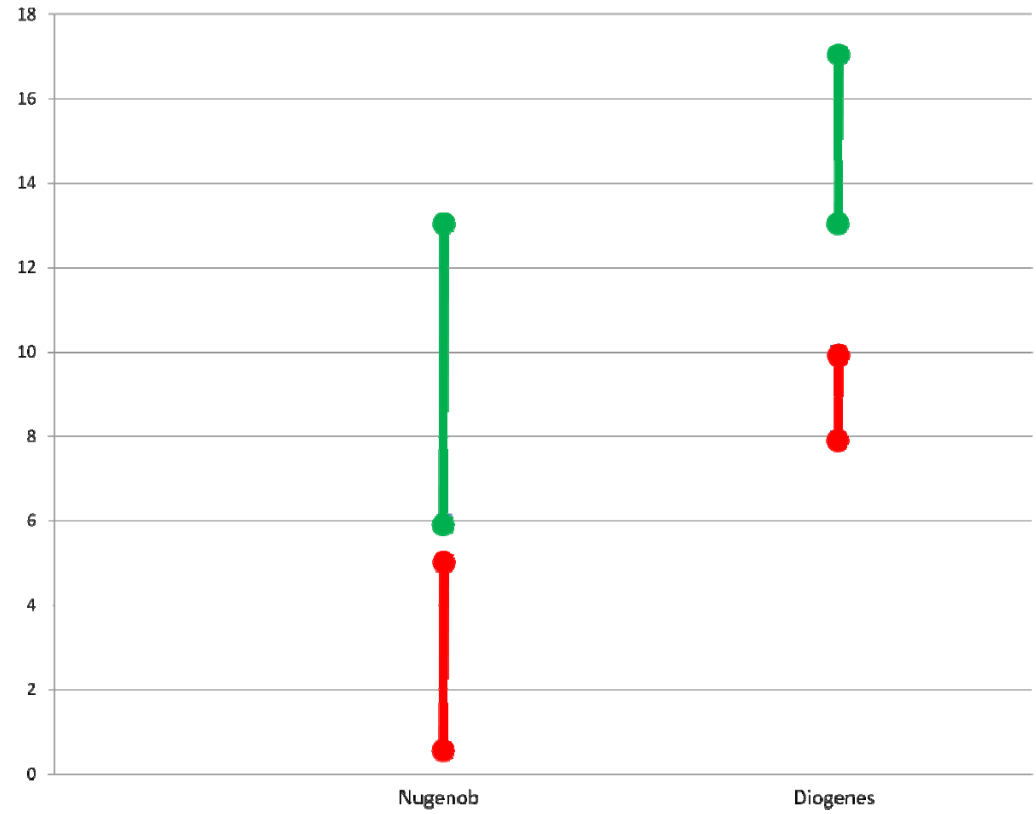


Figure 29 : Distribution de la variation de la perte de poids (en %) chez les sujets de l'étude Nugenob et Diogenes

La durée du régime diffère entre les deux analyses et pour déterminer l'impact de la durée sur le régime il faudrait avoir les données de l'évolution de poids dans les deux analyses à 8 et 10 semaines et comparer les résultats. Mais lors de la rédaction de ce manuscrit, les données ne sont pas encore disponibles.

Comprendre parfaitement la différence entre les performances de prédiction entre les deux analyses que nous avons présentées reste difficile au vu des éléments en notre possession. Néanmoins, nous avons écarté quelques hypothèses pouvant être la cause de cette différence.

L'analyse différentielle conforte nos résultats pour les 2 analyses. Elle montre que les résultats dans Diogenes sont prometteurs et suggère une validation si possible de ces modèles.

Une nouvelle série de puces à ADN est en cours de préparation et qui permettra de déterminer cette fois-ci si on peut prédire la stabilisation de poids après 2mois pour les personnes qui ont réussi leur régime de 8 semaines. Pour cette analyse les données puces de deux groupes de 20 patients chacun va être utilisé pour les modèles prédictifs : le premier groupe de patients a gardé un poids presque stable (10% dans leur poids après deux mois)et le deuxième groupe de patients a repris du poids (entre 50 et 100% de leur poids après deux mois).

4.4 Le cadre de la chirurgie de l'obésité

L'incidence de l'obésité dans les pays occidentaux est sans cesse croissante. Cette « épidémie » est responsable d'une diminution de l'espérance de vie dans la population atteinte, par l'association à l'obésité de nombreux facteurs de comorbidité. De nombreuses modalités thérapeutiques ont été proposées parmi lesquelles la chirurgie bariatrique qui est actuellement le traitement le plus efficace sur le long terme (Sjostrom, Lindroos et al. 2004 ; Maggard, Shugarman et al. 2005)

Cette chirurgie ne vise pas seulement à obtenir une perte de poids, elle vise surtout à contrôler les différentes comorbidités associées à l'obésité (atteintes cardiovasculaires, troubles respiratoires, syndrome d'apnée du sommeil, diabète, problèmes ostéo-articulaires) y compris le déséquilibre hormonal chez la femme, et l'incontinence urinaire (Sjostrom, Lindroos et al. 2004; Raz, Eldor et al. 2005). La chirurgie apporte aussi une amélioration significative de la qualité de vie de ces malades, sur le plan familial, relationnel, et professionnel (Scott, Villegas et al. 2003).

4.4.1 **La chirurgie comme traitement de l'obésité massive**

La chirurgie s'est développée selon deux grands principes : la réduction de la poche gastrique ou gastroplastie à l'origine d'une diminution des apports alimentaires ; et la création d'une insuffisance d'assimilation par court circuit digestif à l'origine d'une malabsorption.

Toutes les procédures ont été réalisées dès 1970 aux USA en chirurgie ouverte et l'expérience américaine considérable peut faire état de plusieurs milliers de malades suivis pendant plus de 20 ans. La révolution est survenue par l'avènement de la laparoscopie (Fig. 2), chirurgie minimale invasive se pratiquant à l'aide d'une micro-caméra et d'instruments fins insérés dans des trocarts passés par ponction dans la paroi.Les avantages sont considérables liés à la diminution des douleurs, à la reprise rapide de la marche et des activités, au raccourcissement du séjour hospitalier. Les indications pour le choix de ce type d'opération correspondent aux recommandations des sociétés scientifiques :

- BMI = poids (kg)/taille (m^2) supérieur à 40 ou bien BMI entre 30-40 lorsqu'il existe une pathologie associée pouvant être améliorée par la perte de poids : HTA, diabète, arthrose radiologiquement prouvée, apnée du sommeil.
- Être âgé de plus de 18 ans et avoir une obésité stable depuis plus de 5ans.
- Échec de régimes alimentaires ou médicamenteux depuis plus d'un an.
- Absence de pathologie endocrinienne

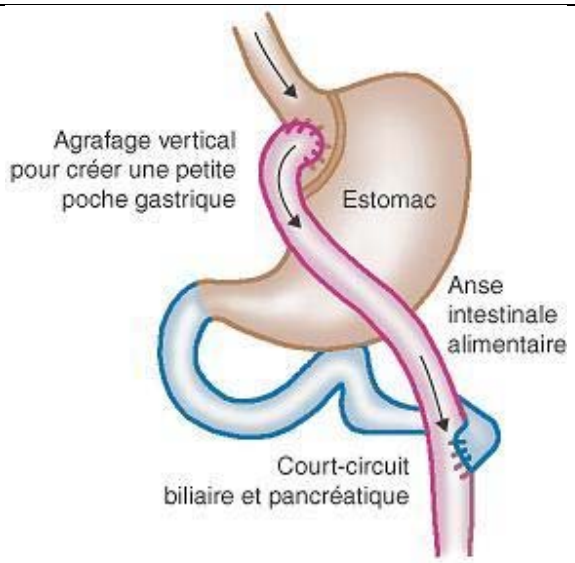


Figure 30 : Schéma du bypass gastrique

L'intervention chirurgicale n'est que l'un des éléments du programme dans lequel entre le patient qui a pris la décision de se faire opérer. Dans le cas de l'obésité morbide, tout au long des étapes de sa vie future, le patient entre dans un programme où il est suivi régulièrement par l'équipe du centre de l'obésité sous la direction du médecin. Le suivi est essentiel pour dépister éventuellement des complications, apprécier la courbe d'amaigrissement, vérifier l'absence de carences alimentaires, contrôler l'amélioration des maladies associées.

La mise en place d'un système d'aide à la décision qui permet de guider les médecins dans leur choix concernant la prescription d'une chirurgie de type bypass peut s'avérer d'une grande utilité pour les patients du fait de la lourdeur de l'opération et du suivie demandé.

4.4.2 **Prédiction de la perte de poids suite à un Bypass**

La prédiction de la perte de poids suite à une intervention chirurgicale reste un champ de recherche peu exploré. Larsen et al (Larsen, Geenen et al. 2004) ont essayé d'utiliser la personnalité comme prédicteur du maintien du poids suite une chirurgie bariatrique et dans le cas d'une obésité morbide et ceci à court terme mais aussi à long terme. Leurs travaux effectués

sur 168 sujets (143 femmes et 25 hommes, âgé entre 18 et 58 ans avec un âge moyen de 37 ans et ayant une IMC préopératoire de 45.9 $\pm$ 5.6 Kg/m²) et ayant répondu à un questionnaire de personnalité un an et demi avant l'opération montre que les paramètres de la personnalité ne permettent pas de prédire la perte de poids à court terme il en est de même pour la prédiction à long terme. Lee et al (Lee, Lee et al. 2007) quant à eux, appliquent des techniques de data mining pour prédire la perte de poids après une chirurgie bariatrique (de type bypass ou anneau) à partir des données biocliniques. Leur étude comporte 249 patients (177 femmes et 72 hommes, avec un âge moyen de 33 $\pm$ 9 ans) ; parmi eux, 208 (83.5%) avait un résultat positif après l'opération et 41 (16.5%) avait un résultat négatif. La méthode de la régression logistique a été appliquée ainsi que les réseaux de neurones. Les auteurs montrent que la variable prédictive avec la régression logistique est le type d'opération. Avec les réseaux de neurones, les variables prédictives sont dans l'ordre d'importance l'hémoglobine glycosylée (HbA1c) qui permet de déterminer la concentration du glucose dans les 3 précédents la prise de sang, les triglycérides et le type d'opération. Nous avons voulu étendre ce travail exploratoire en analysant la base 'bypass' de l'hôpital « Hôtel Dieu » afin de retrouver d'une part d'autres variables biocliniques pertinentes et rechercher aussi des profils de patients pouvant expliquer la réponse à une opération de type bypass.

4.4.2.1 **Etude exploratoire de la base de donnée « bypass »**

Pour cette analyse nous disposons d'une base de données clinique de la chirurgie gastrique qui recense les paramètres biocliniques avant l'intervention et aussi après celle-ci à des intervalles réguliers pour tous les patients. L'objectif est de mettre en place un système d'aide à la décision qui aiderait les cliniciens dans leur choix thérapeutique et ainsi permettre une meilleure prise en charge des patients. Pour la réalisation du modèle d'apprentissage nous avons à disposition la base de données de la chirurgie gastrique que nous avons présentée dans 2.5.1.2.

Pour cette étude exploratoire, nous nous sommes intéressés à la contribution des arbres de décision qui donne une information pratique sur les facteurs liés a la perte de poids après un bypass à 3 mois et nous avons utilisé l'outil SPSS Clémentine pour réaliser ce travail.

Nous avons utilisé la variation de l'IMC qui est un indicateur de la masse corporelle. Nous obtenons l'arbre de décision de la Figure 31.

Ce modèle met en avant les variables suivantes :

- L'âge
- L'IMC préopératoire
- La CRP préopératoire
- La GGT préopératoire
- Le sexe préopératoire
- La SAA préopératoire

Nous avons estimé la précision de ce modèle en validation croisée et nous avons obtenu 57,7 ± 4,9%. Suite à ces résultats, nous avons essayé d'appliquer des méthodes qui nous permettraient d'améliorer les performances et aussi de retrouver des profils de patients. Nous avons vu dans la littérature plusieurs propositions de combinaison de modèles permettant d'avoir une meilleure vision des problèmes complexes et nous avons voulu explorer cette direction avec les données de la chirurgie qui sont en notre possession.

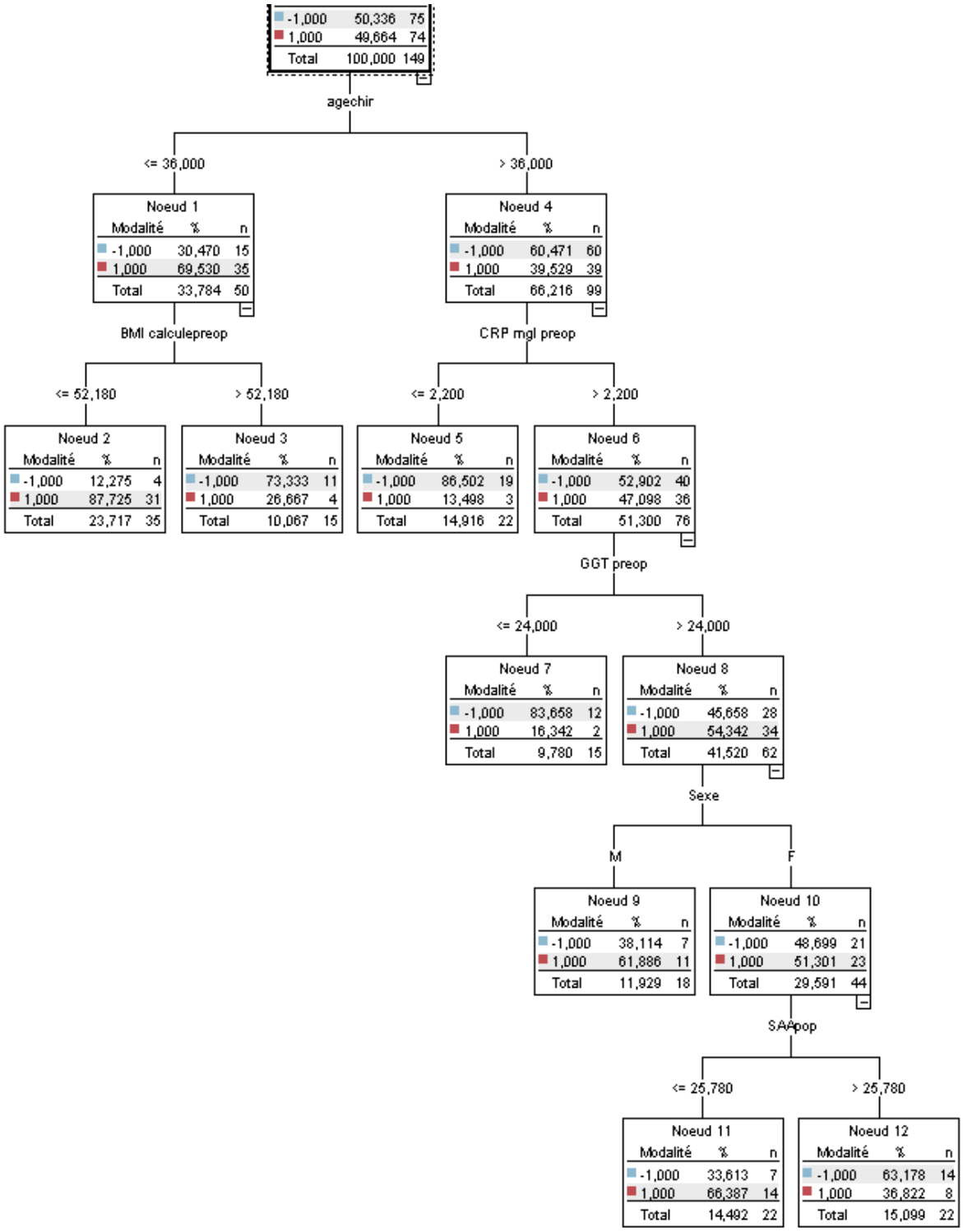


Figure 31: Arbre de décision de la variation de l'IMC après 3 mois d'un Bypass

Cet arbre de décision est équivalent aux règles suivantes :

- Si l'âge est inférieur ou égale à 36 ans et l'IMC est supérieur 52.18 alors le sujet appartient à la classe (-, c.à.d. mauvais répondeur)
- Si l'âge est inférieur ou égale à 36 ans et l'IMC est inferieure ou égale à 52.18 alors le sujet appartient à la classe (+, c.à.d. bon répondeur)
- Si l'âge est supérieur à 36 ans et la CRP est inférieure ou égale à 2.2 alors le sujet appartient à la classe (-, c.à.d. mauvais répondeur)
- Si l'âge est supérieur à 36 ans et la CRP est supérieure à et la GGT est inférieure ou égale à 24 alors le sujet appartient à la classe (-, c.à.d. mauvais répondeur)
- Si l'âge est supérieur à 36 ans et la CRP est supérieure à et la GGT est supérieure à 24 et c'est un homme alors le sujet appartient à la classe (+, c.à.d. bon répondeur)
- Si l'âge est supérieur à 36 ans et la CRP est supérieure à et la GGT est supérieure à 24 et c'est une femme et la SAA est inférieure ou égale à 25.78 alors le sujet appartient à la classe (+, c.à.d. bon répondeur)
- Si l'âge est supérieur à 36 ans et la CRP est supérieure à et la GGT est supérieure à 24 et c'est une femme et la SAA est supérieure à 25.78 alors le sujet appartient à la classe (-, c.à.d. mauvais répondeur)

4.4.2.2 **Décomposition de problème complexes en sous problème simples.**

Parmi les problèmes d'apprentissage, il existe une catégorie de problèmes qui sont appelés dans la littérature : difficiles, complexes, hétérogènes. Ces problèmes ont souvent la particularité d'avoir des résultats non satisfaisants, quelque soit la méthode standard utilisée (arbre de décision, SVM, MLP, ...). En d'autres termes, ces méthodes obtiennent légèrement un meilleur résultat qu'un algorithme qui ferait un tirage aléatoire. Dans d'autres problèmes, appelés aussi problèmes difficiles, on dispose de données issues de sources différentes ou des données mixtes. Ces problèmes existent dans de nombreux domaines; souvent ils sont mal posés, les bases d'apprentissage disposent de peu d'observations et/ou le nombre de variables est trop grand. Parmi ces domaines, on retrouve le domaine médical, spécialement les bases cliniques ou les bases biopuces (Tamayo, Slonim et al. 1999; Pavlidis, Weston et al. 2001). De nombreuses méthodes sont développées pour ce genre de problème qui consiste à utiliser les techniques de sélection de variables (Golub, Slonim et al. 1999; Xing, Jordan et al. 2001; Long and Ding 2005), la reformulation de problème en utilisant le classement heuristique, (Clancey 1985), et le boosting (Long and Vega 2003). D'autres méthodes consistent à combiner ou fusionner les classeurs (Egmont-Petersen, Dassen et al. 1999; Liu and Yuan 2001).
Dans ce chapitre, nous présentons un modèle combinant deux modèles d'apprentissage pour aborder les problèmes de classement difficiles ou "complexes".

La question qui se pose pour ce genre de base complexe, *faut-il aborder le problème de classement d'une manière globale ou trouver un moyen de le diviser en sous-problèmes?* Le modèle consiste à partitionner les données pour sélectionner le classeur adéquat pour chaque classe de partition (Liu and Yuan 2001).

Quelques méthodes spécifiques combinant l'apprentissage non supervisé et supervisé, aussi bien dans le domaine de la classification hiérarchique et les arbres de décision que pour la recherche de partitions ont été développées. Dans le domaine de la classification hiérarchique, (Kim, Pang et al. 2003; Benabdeslem 2006) proposent des modèles combinant la classification hiérarchique puis entraînent un SVM binaire dans chaque nœud du dendogramme de la classification hiérarchique (Sungmoon Cheong and Lee 2004; Pang, Kim et al. 2005) proposent d'utiliser une autre méthode de partitionnement qui permet de construire une partition de sous

ensembles ordonnés sous forme d'arbre binaire afin d'apprendre un SVM binaire au niveau de chaque nœud de l'arbre.

Dans (Wei, Xin et al. 2004), les auteurs proposent d'utiliser les cartes topologiques de Kohonen (Kohonen 1998) pour filtrer les données. Les observations non étiquetées héritent de l'étiquette de la classe du vote majoritaire de son sous-ensemble. A la fin de cette phase, un seul SVM est appris sur l'ensemble d'apprentissage initial ré-étiqueté, sans prendre en compte la partition de données. Nous trouvons aussi l'utilisation d'autres méthodes de partitionnement comme le k-means lorsqu'un sous-ensemble de la partition est constitué à 100% d'une seule classe alors toutes les observations de ce sous-ensemble sont remplacées par leur référent (représentant), calculé par le K-means (K-moyennes), dans la base d'apprentissage dédiée au SVM. Ce procédé permet de réduire significativement la taille de la base et par conséquent le temps de calcul sur de grandes bases de données. D'autres méthodes sont aussi inspirées des méthodes de partitionnement et de classement comme la définition de cartes topologiques dans l'espace de re-description (Sungmoon Cheong and Lee 2004) ou l'utilisation des vecteurs supports pour définir une partition (Ben-Hur, Horn et al. 2002).

Toutes ces méthodes ont un point commun, celui de montrer que la visualisation et le prétraitement des données sont des étapes importantes dans la phase exploratoire de l'analyse de données. Cette phase permet d'inclure les connaissances de l'expert du domaine avant la phase de classement. La difficulté en classement augmente d'autant plus qu'il s'agit de données complexes, par exemple données mixtes (quantitatives et qualitatives) ou des données biomédicales, pour lesquelles il existe moins de méthodes standards.

L'approche proposée consiste à diviser le problème global de classement en sous-problème de classement guidé par la structure et l'organisation des données de la base dans l'espace des données. Ce modèle est basé sur le partitionnement des données, avec une méthode non supervisée, en une partition constituée de plusieurs sous-ensembles organisés, en tenant compte de la typologie des données, qui vont servir à définir un classeur pour chacun en utilisant les SVMs (Vapnik 1995; Schèolkopf, Burges et al. 1999). La tâche de partitionnement des données de notre modèle est réalisée à l'aide des cartes auto-organisatrices (Kohonen 1998).

Les cartes topologiques sont utilisées dans notre modèle parce qu'elles permettent à la fois d'être utilisée comme outil de visualisation et de partitionnement non supervisé de différents types de données (quantitatives et qualitatives). Elles permettent de projeter les données sur des espaces discrets qui sont généralement de dimensions deux. Le modèle de base, proposé par Kohonen(Kohonen 1998), est uniquement dédié aux données numériques. Des extensions et des reformulations du modèle de Kohonen ont été proposées dans la littérature (Bishop, Svensen et al. 1998; Lebbah, Chabanon et al. 2002; Lebbah, Chazottes et al. 2005).Une généralisation des cartes topologiques sera présentée dans ce chapitre.

Les machines à vecteurs de support ont été utilisées dans notre modèle parce qu'elles s'avèrent particulièrement efficaces, car elles peuvent traiter des problèmes mettant en jeu un grand nombre de variables ou un petit nombre d'observations (individus), et qu'elles assurent une solution unique (pas de problèmes de minimum local comme pour les réseaux de neurones). L'algorithme sous sa forme initiale revient à chercher une frontière de décision linéaire entre deux classes, mais ce modèle peut considérablement être enrichi en se projetant dans un autre espace permettant ainsi d'augmenter la séparabilité des données. Ce cas d'utilisation des machines à vecteurs de support est le plus utilisé, car la plupart des problèmes réels sont non linéairement séparables.

4.4.2.3 **Méthode Hybride : CT-SVM**

Pour certains problèmes de classement, il est préférable de décomposer le problème global de classement en sous-problèmes pour améliorer les performances en classement, (Platt 1999; Gama and Brazdil 2000; Kuncheva 2002). Par exemple, si l'on dispose d'une base de données où certaines observations sont linéairement séparables et les autres sont non linéairement séparables, alors il est possible de décomposer la base entière en deux sous-ensembles et d'entraîner un classeur SVM pour chacun des sous-ensembles. Ce cas d'utilisation des machines à vecteurs de support dans le cas non linéaire est le plus intéressant car la plupart des problèmes réels sont non linéairement séparables. Il est évident que la détermination du nombre d'observations et par conséquent la taille de la partition utilisée pour l'apprentissage de chaque SVM est important d'un point de vue de la théorie de l'apprentissage (Vapnik 1995). Dans cette section, nous ne présentons pas un indice permettant d'estimer la taille de la

partition, mais nous allons présenter par la suite un modèle de classement qui permet d'augmenter les performances en classement en utilisant le partitionnement des observations.

Dans (Kuncheva 2004), l'auteur fournit une démonstration pour ce type de modèle. Si l'on suppose que l'on dispose de **S** classeurs notés **Cla$_i$** associés à différents sous-ensembles P_i et si on note par $p(Cla_i \mid P_i)$ la probabilité du classement correct par le classeur Cla_i dans le sous-ensemble P_i, alors la densité de probabilité du classement correct de notre système de partitionnement et de classement s'écrit:

$$p(correct) = \sum_{i=1}^{S} p(P_i) p(Cla_i \mid P_i)$$

où $p(P_i)$ est la probabilité a priori que l'observation soit générée dans le sous-ensemble P_i. Pour maximiser ce mélange de probabilité, on choisit $p(Cla_i / P_i)$ tel que $p(Cla_i \mid P_i) \geq p(Cla_j \mid P_j), j = 1..S$.

Afin de simplifier le problème de classement, notre approche consiste à entraîner des SVMs ($Cla_i = SVM$) différents avec chaque sous-ensemble d'une partition P de la base A. Ceci permet de redéfinir des espaces de redescription différents (ou les mêmes) pour chaque sous-ensemble $P_c \in P$, L'objectif de notre modèle CT-SVM est d'améliorer la discrimination en entraînant un SVM pour chaque sous-ensemble $P_c \in P$ qui a plus d'une classe. Pour les sous-ensembles, qui sont composés d'observations de la même classe, aucun SVM ne sera entraîné. L'algorithme des cartes topologiques mixtes est utilisé pour définir une partition de la base d'apprentissage.

Afin de réduire la partition et par conséquent le nombre de SVMs entraînés, nous avons utilisé la classification hiérarchique (CAH), sur l'ensemble des référents W de la carte pour réduire la partition ainsi le nombre de sous-ensembles (Ripley and Hjort 1995; Yacoub, Badran et al. 2001). Cette phase de réduction de la partition, qui consiste à fusionner certains sous-ensembles, est optionnelle et, elle peut être déterminée en interaction avec les experts et après visualisation des cartes topologiques.

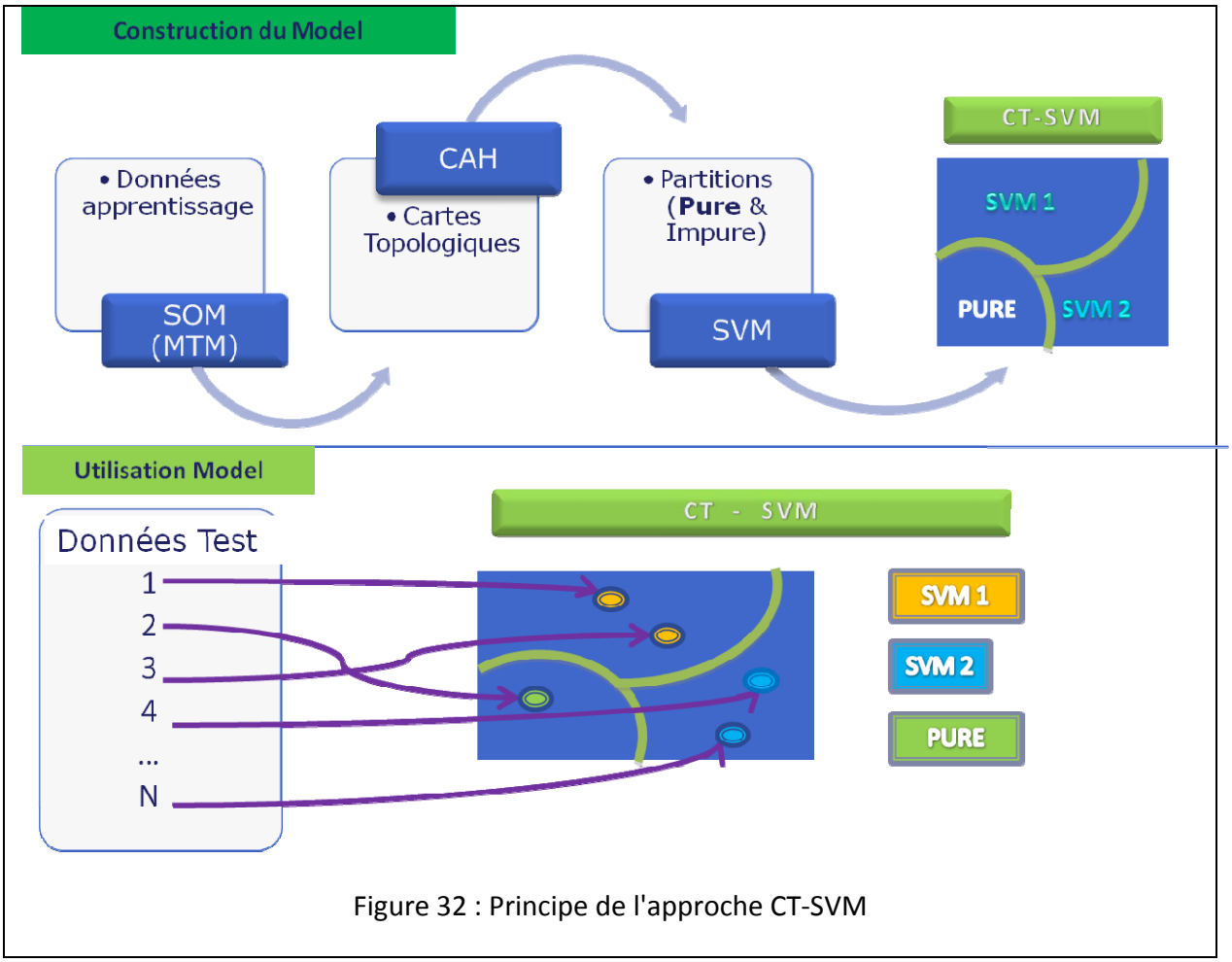


Figure 32 : Principe de l'approche CT-SVM

L'algorithme de note modèle CT-SVM est le suivant:

Pour un nombre de sous-ensembles S fixé faire:

- **Phase 1:** Construction d'une partition $P=\{P_1,...,P_{N_{cell}}$ / en utilisant les cartes topologiques avec l'algorithme défini dans la section 2.1.
- **Phase 2 (optionnelle):** Si $S < N_{cell}$ appliquer la classification hiérarchique (CAH) pour construire la nouvelle partition $P=\{P_1,...,P_S / 1 \le S \le N_{cell}\}$
- **Phase 3:** Détecter l'ensemble des indices I_p des sous-ensembles purs tel que $I_p=\{c / \forall \mathbf{z} \in P_c, \chi(\varphi(\mathbf{z}))=c, vote(P_c)=y_c\}$. y_c est l'étiquette du vote majoritaire à 100% du sous-ensemble P_c.
- **Phase 4:** Apprentissage du SVM pour chaque sous-ensemble P_i tel que $i \notin I_p$.

Pour l'apprentissage des cartes topologiques mixtes, nous avons utilisé un programme que nous avons déjà développé en *C++*. Nous avons aussi utilisé les programmes et l'heuristique développée par l'équipe de Kohonen (Vesanto, Himberg et al. 1999) pour estimer la dimension de la carte. Pour l'apprentissage du modèle SVM dans le cas du multi-classe, nous avons utilisé le modèle DAG-SVM (Directed Acyclic Graph SVM) développé par (Vesanto and Alhoniemi 2000).

Avec ce modèle CT-SVM, la topologie ou la forme des observations est présentée par les cartes topologiques. Lorsqu'on présente une nouvelle observation qui n'a pas participé à la phase d'apprentissage, elle sera projetée d'abord sur la carte topologique avec la fonction d'affectation associée φ, puis on utilisera la fonction d'affectation χ, pour sélectionner le sous-ensemble qui va déterminer le classeur SVM associé. Cette méthode d'affectation de notre classement permet de comprendre le comportement d'une observation à travers son référent $\mathbf{w}_c$. Si on note par svm_r la fonction de classement du modèle SVM du sous-ensemble P_r alors la fonction d'affectation globale de notre système s'écrit comme suite:

$$y_i = \begin{cases} svm_{\chi(\varphi(\mathbf{z}_i))} & \text{si } \chi(\varphi(\mathbf{z}_i)) \notin I_p \\ vote(P_{\chi(\varphi(\mathbf{z}_i))}) & \text{sinon} \end{cases},$$

où I_p est l'ensemble des indice des sous-ensembles pures. $\chi(c) = c$ si $P = \{P_1,...,P_c,...,P_{N_{cell}}\}$ et $\chi(c) = 1$ si $P = A$.

4.4.2.4 **Résultats**

Nous avons évalué l'apport de la combinaison des cartes topologiques et des méthodes d'apprentissage supervisé (SVM, arbre de décision et forêt aléatoires) à l'amélioration de la prédiction par rapport à l'utilisation des méthodes d'apprentissages supervisés seules.

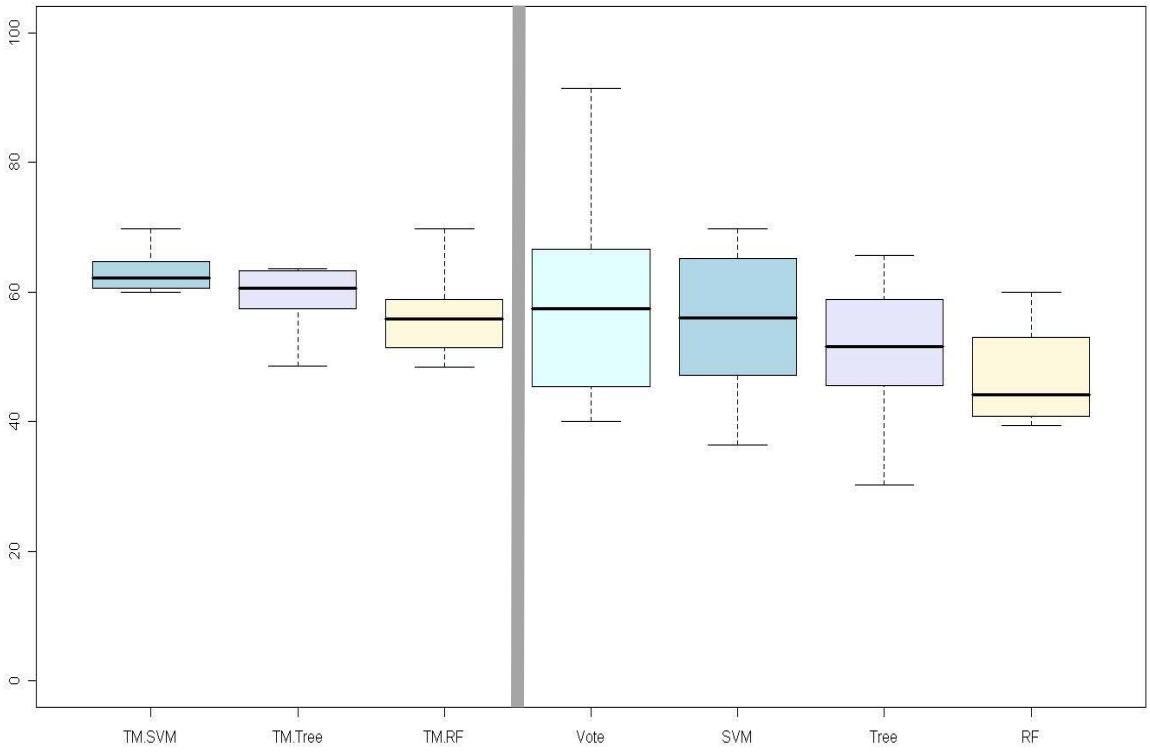


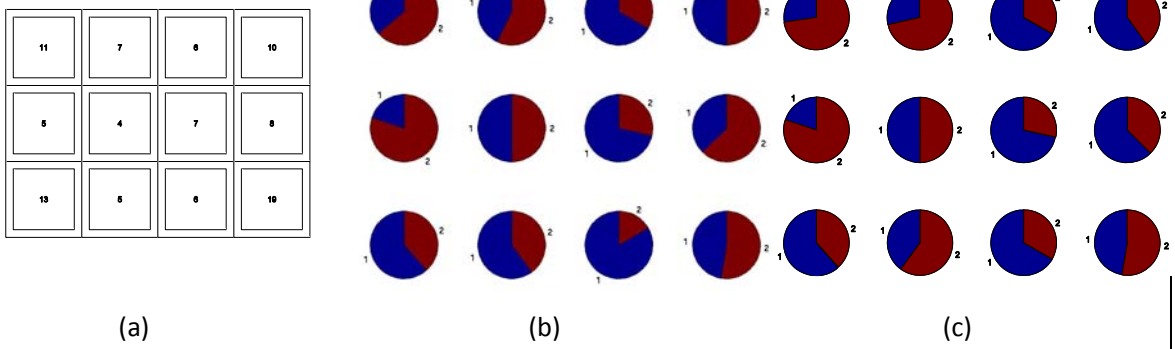
Figure 33 : Comparatif des méthodes de classification avec et sans le partitionnement par les cartes topologiques

La Figure 33 montre que la décomposition du problème et l'application des cartes améliorent légèrement la précision de la prédiction et ceci avec les trois méthodes utilisées. Cette application des cartes topologique réduit aussi la variance dans les résultats. Nous avons les meilleurs résultats avec les machines à vecteur de support. Nous nous sommes intéressés par la suite à l'analyse de profils pour essayer de déterminer des caractéristiques communes a des groupes de patients qui permettraient d'avoir des pistes biologiques capables d'éclaircir certains mécanismes de la perte de poids.

4.4.2.5 **Analyse des profils**

Puisque notre modèle utilise les cartes topologiques, on dispose d'un pouvoir de visualisation de la partition. L'application d'abord des cartes topologiques mixtes, va nous permettre d'analyser la répartition des observations et par conséquent les sous-ensembles qui ont servi au classement. L'apprentissage d'une carte de dimension 3×4 cellules effectué sur la

base entière des patients, avec l'hyper-paramètre F=0.01, fournit pour chaque cellule un référent $\mathbf{w}_c$ composé de deux parties: la partie quantitative $\mathbf{w}_c^r$ et la partie qualitative $\mathbf{w}_c^b$ codée avec le codage disjonctif binaire.



(a) (b) (c)

Figure 34 : Cartes topologiques 3×4 ($P=\{P_1,P_2,...,P_{12}\}$). (a) Cardinalité des sous-ensembles (b) et (C) Répartition des pertes de poids respectivement à 3 mois et à 6 mois. . 1: Pas de perte de poids; 2: perte de poids.

La Figure 34.a présente la répartition des observations. On observe que la partition obtenue a permis de bien distribuer les observations sur 12 cellules de l'ensemble de la partition $P=\{P_1,...,P_{12}\}$. La figure 8.b présente la même répartition en distinguant ceux qui ont perdu ou non du poids à 3 mois par rapport à la médiane de l'ensemble des patients. La figure 3.c présente la même répartition de perte de poids à 6 mois. On constate que les sous-ensembles sont mélangés.

A l'aide de cette carte topologique 3×4, il est possible d'effectuer un certain nombre d'analyses de la base étudiée. Notre premier objectif est celui de partitionner les données, en prenant en compte leurs spécificités (données mixtes) pour augmenter les performances en classement. En plus du classement, il est possible d'utiliser le pouvoir de visualisation des cartes topologiques. Pour visualiser la carte topologique, nous nous sommes limités à analyser les effets dus à quelques variables pour lesquels l'exactitude des propriétés médicales retrouvées peuvent êtres vérifiées. En regardant à la fois les trois Figure 34.a, Figure 34.b et Figure 34.c, le médecin a détecté globalement trois grands groupes. Pour s'approcher de la partition du médecin, nous avons appliqué la CAH avec les référents de la carte pour avoir 4 sous-

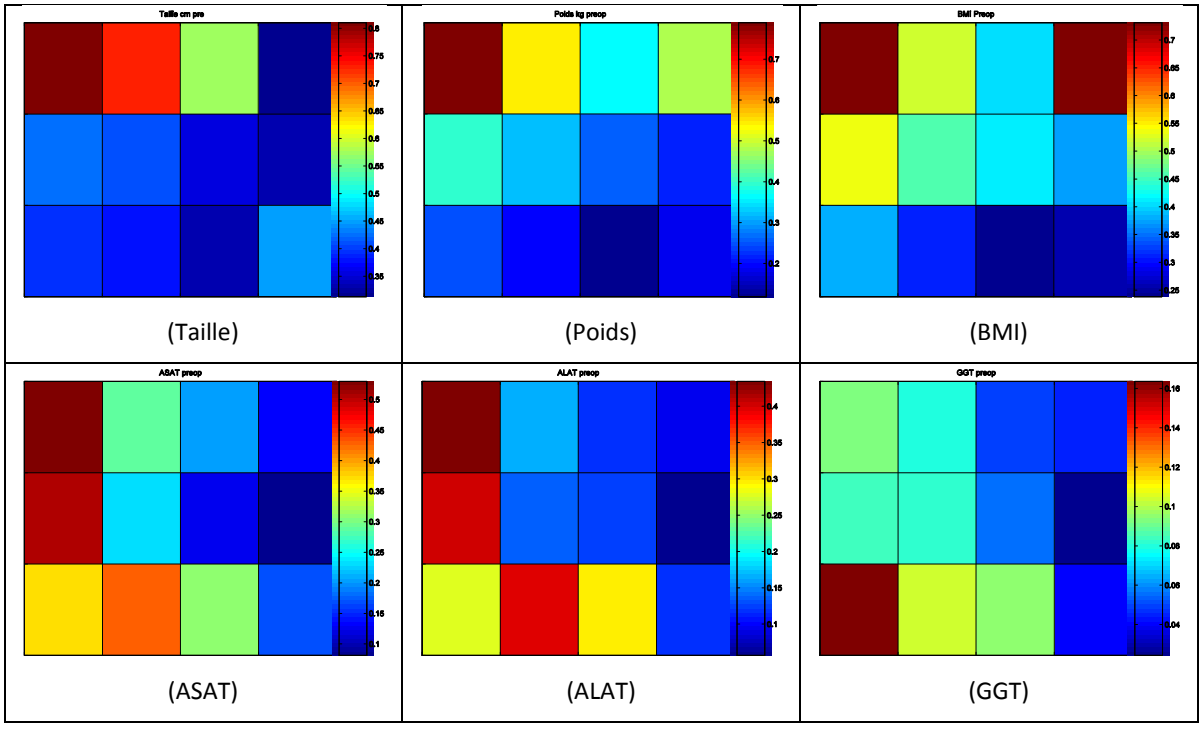
ensembles, $P = \{P_1, P_2, P_3, P_4\}$. La figure 8 présente la partition avec 4 sous-ensembles numérotés de 1 à 4. Cette répartition des données en quatre sous-ensembles et la répartition du médecin en trois sous-ensembles correspondent à la taille de la partition utilisée dans la phase de la validation croisée décrite ci-dessous. En visualisant à la fois les figures 3,4, 5, 6, 7 et la figure 8, il est possible de demander au médecin de définir des profils de patients. Ces profils vont servir à décrire les paramètres (variables) liés à la perte de poids et fournir des hypothèses de travail sur la résistance à la perte de poids fourni par le classifieur.

Trois grands profils de patients sont définis selon la cinétique de perte de poids à 3 mois et à 6 mois. Le profil 1 est plutôt un bon profil par rapport aux pertes de poids à trois mois (figure 3.b) et 6 mois (figure 3.c) et correspond aux deux sous-ensembles P_1 et P_2 de la CAH. Le profil 2 est caractérisé par une perte de poids moyenne à 3 mois et 6 mois et correspond approximativement au sous ensemble P_4 de la CAH. Enfin, le profil 3 est caractérisé par une perte de poids médiocre à 3 mois dont l'amplitude diminue à 6 mois, ce qui aboutit à dénommer ce profil comme un "mauvais" profil en terme de perte de poids. Ce profil correspond au sous-ensemble P_3 de la CAH. Nous détaillons par la suite les deux profils 1 et 3 par rapport aux différentes variables clinico-biologiques.

Le profil 1 est caractérisé par un poids, un IMC et une Dépense Energétique de Repos mesurée par calorimétrie (DERm) élevés. Les patients appartenant à ce profil ont une glycémie à jeun et insulinémie élevées sans être diabétiques. Il s'agit donc de patients insulinorésistants avant le stade de diabète. Le reste du profil métabolique est caractérisé par des HDL plutôt bas, des triglycérides (TG) et enzymes hépatiques (ASAT, ALAT et GGT) élevés. Dans les classes qualitatives "HTA" (hypertension) ou "SAS" (Syndrome d'apnées du sommeil) ces patients sont classés "oui". D'un point de vue inflammatoire, la CRP, la férritinémie (FERR), la SAA et l'orosomucoide (ORO), toutes des protéines de la phase aiguë de l'inflammation, sont modérément élevées. Sur le plan nutritionnel, la TSH est basse, le profil protéique (albumine, préalbumine, RBP) et vitaminique est favorable, sans déficit. En conclusion pour ce profil, il s'agit de patients avec un poids très élevé, mais dont le profil métabolique n'est pas trop évolué (sans diabète), sans inflammation importante et un bon profil nutritionnel.

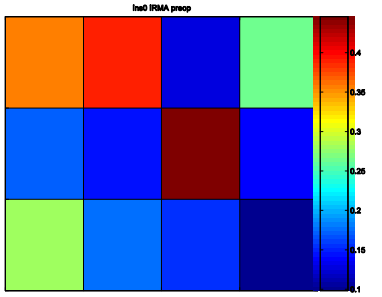
Le profil 2 correspond à des patients ayant un BMI élevé et une leptine élevée. Ils sont insulinorésistants, mais pas diabétiques. Ils ont majoritairement une HTA et un SAS. Les paramètres hépatiques et métaboliques sont normaux. L'adiponectinémie (ADIPO) est plutôt basse. En revanche, les paramètres inflammatoires (SAA et CRP) sont très élevés. Sur le plan nutritionnel, la TSH est normale haute et les marqueurs nutritionnels sont bas (bilan protéique avec albumine, préalbumine et RBP, fer, vitamines A, E, B1, B12). Le profil 3 est un profil intermédiaire en terme de paramètres clinico-biologiques.

En conclusion, les deux profils de patients 1 et 2 sont caractérisés par des paramètres clinico-biologiques différents, notamment en terme de marqueurs d'inflammation et nutritionnels et sont aussi différents en termes de profil de perte de poids à 3 mois et 6 mois. Nous pouvons donc formuler l'hypothèse que le statut nutritionnel et l'état d'inflammation des patients avant chirurgie pourraient être des éléments liés à la résistance à la perte de poids.

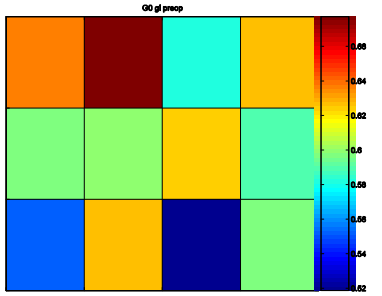


(Taille) (Poids) (BMI)

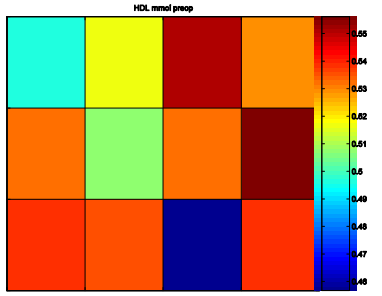
(ASAT) (ALAT) (GGT)



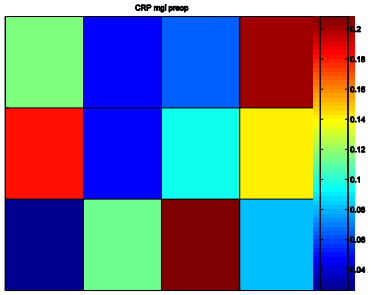
(INS)



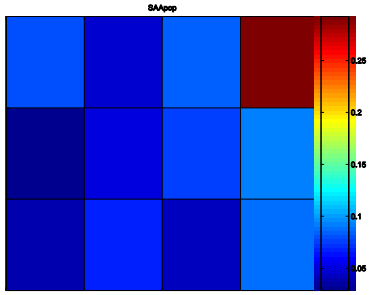
(GLY)



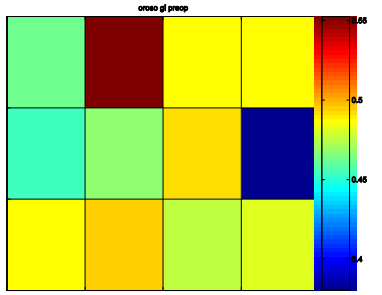
(HDL)



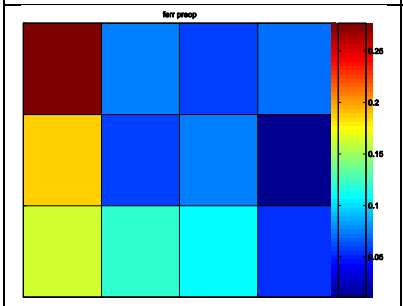
(CRP)



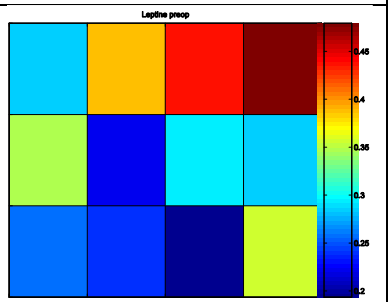
(SAA)



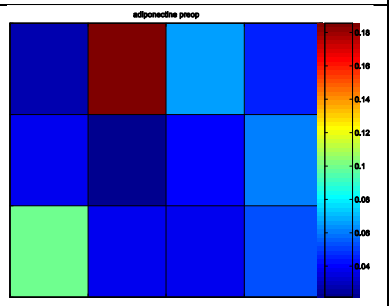
(ORO)



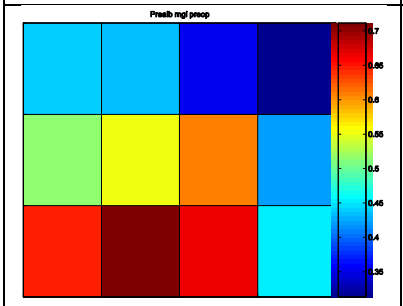
(FERR)



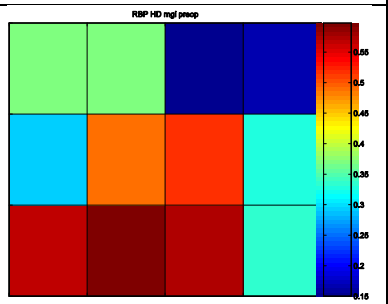
(LEP)



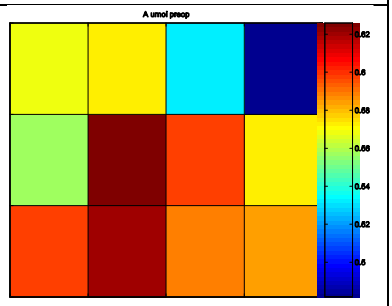
(ADIPO)



(PREALB)



(RBP)



(A)

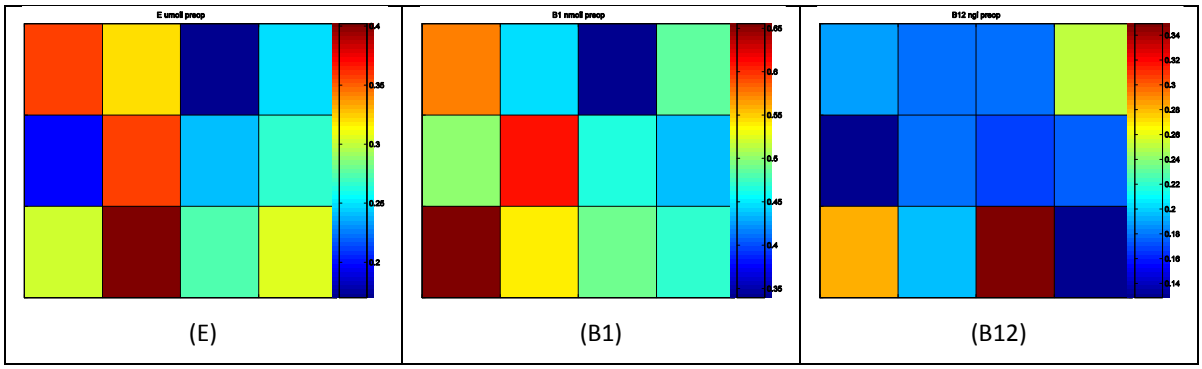


Figure 35 : Cartes topologiques décrivant la variation sur les variables Taille,poids,BMI (Body Mass Index), ALAT, ASAT,GGT, INS(insuline), GLY (glycémie),HDL, CRP,SAA,ORO (orosomucoide),FERR (férritinémie),LEP (leptine),ADIPO (adiponectinémie),PREALB (préalbumine),RBP, A, E, B1, B12.

4.4.2.6 **Discussion**

Dans cette partie, nous avons présenté un modèle de classement hybride, associant une méthode de partitionnement et une méthode de classement qui sont respectivement, les cartes topologiques et les SVMs. Ce modèle utilise l'organisation des données fournies par les cartes topologiques mixtes pour subdiviser l'espace des données afin d'apprendre un SVM spécifique pour chaque sous-espace des données. Notre modèle CT-SVM utilise la partition résultat des cartes topologiques, pour associer un SVM à chaque sous-ensemble de la partition avec des hyper-paramètres différents si cela est nécessaire. Ceci permet d'améliorer légèrement la précision et réduire la variance des résultats. Les cartes topologiques permettent aussi de déceler des profils de patients et comprendre leurs comportements métaboliques.

Étant donné que la précision de la prédiction de la perte de poids est limitée, et vu que la base en notre possession contient des données de variation d'autres variables biocliniques nous nous sommes intéressés à l'étude de la prédiction de ces variables qui permettront de prédire une amélioration de l'état de santé d'un patient suite à l'intervention chirurgicale puisque cela aussi utile que de savoir si le patient va perdre du poids ou non.

4.4.3 **Prédiction de l'évolution des paramètres bioclinique suite à un Bypass**

Nous avons essayé dans cette partie d'investiguer les autres variables biocliniques pour lesquelles nous avons des données disponibles à 3 mois et à 6 mois. Nous avons retenu pour cette analyse les variables suivantes :

- IMC: indice de masse corporelle
- TG: triglycéride
- HDL: lipoprotéines de haute densité
- INS: insulinémie
- GLY: glycémie
- QUICKI: indice de sensibilité à l'insuline

En effet la connaissance de l'évolution de ces paramètres à travers le temps est un indicateur de l'état de santé des patients après la chirurgie, et prédire l'amélioration de ces paramètres après une chirurgie est aussi important que la connaissance de la perte de poids et constitue ainsi un critère de décision de la réussite de l'opération. À notre connaissance, il n'existe pas de critères biologiques dans la littérature pour définir un seuil distinguant une bonne d'une mauvaise évolution de ces paramètres, nous avons donc décidé de prendre 3 critères de seuillages pour la construction des classes basées sur la distribution des données telles que l'illustre la Figure 36.

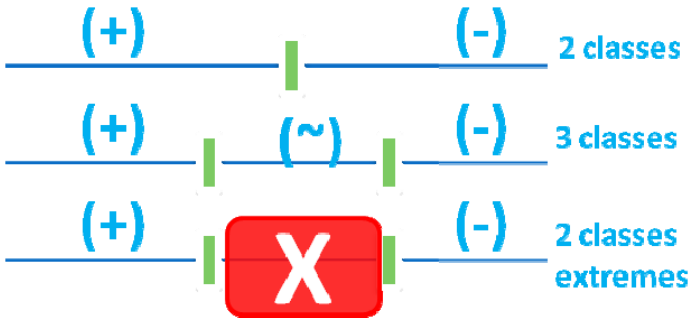


Figure 36 : critères de sélection des classes pour l'évolution des paramètres biocliniques

Le choix des classes a été défini selon les critères suivants :

- 2 classes par la médiane: nous construisons 2 classes, une classe « mauvaise évolution » pour les sujets ayant une variation supérieure à la médiane du groupe et une classe « bonne évolution » pour les sujets ayant une variation inférieure à la médiane du groupe. Ceci pour toutes les variables que nous avons sélectionnées à l'exception de l'HDL où les classes sont inversées puisque l'augmentation de ces lipoprotéines de haute densité a un effet protecteur.
- 3 classes par le 1er et le 3^{ième}quartile : nous construisons dans ce problème 3 classes, une classe « mauvaise évolution » pour les sujets ayant une variation supérieure au 3^{ième} quartile du groupe, une classe « pas d'évolution » pour les sujets ayant une variation entre le 1er et le 3^{ième} quartile du groupe et une classe « bonne évolution » pour les sujets ayant une variation inférieure à la médiane du groupe.
- 2 classes extrêmes : nous ne gardons dans ce cas que les classes « mauvaise évolution » et « bonne évolution » de la deuxième définition.

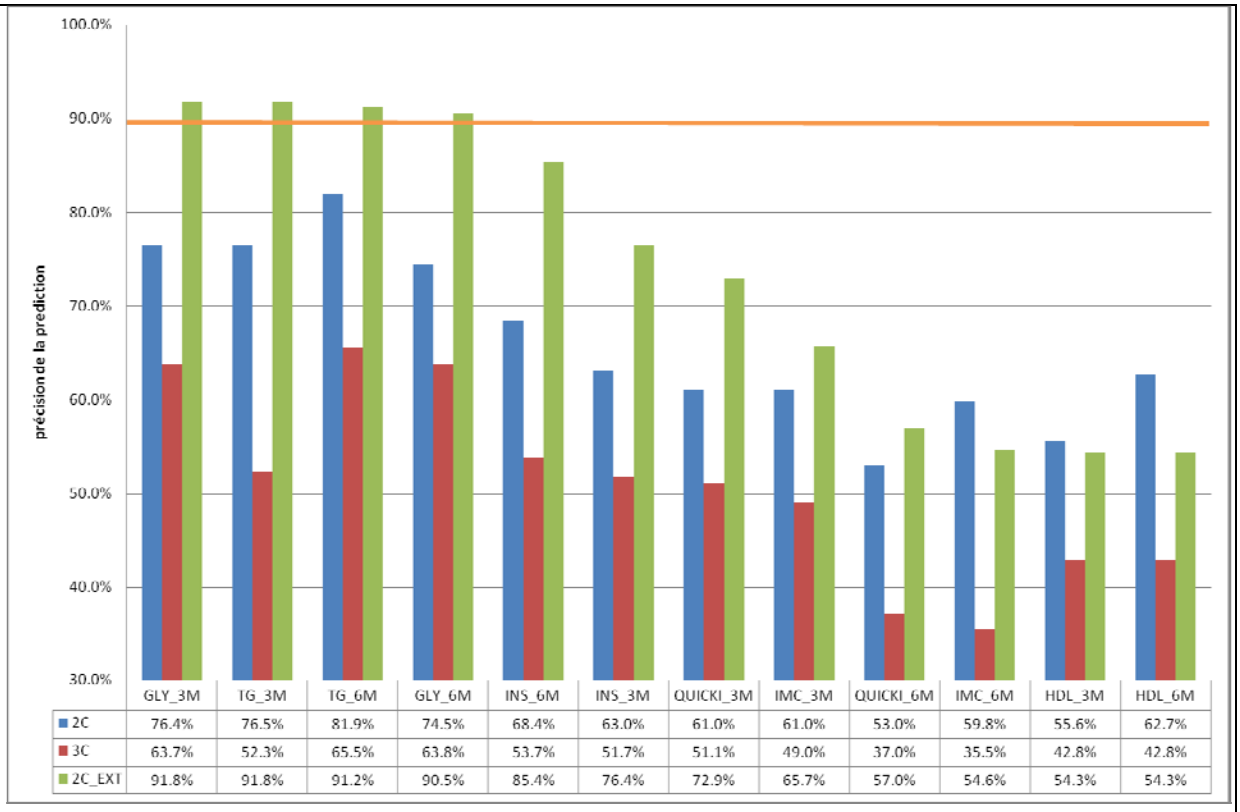
Les valeurs des médianes et du 1er et 3^{ième} quartile des variations à 3 mois et 6 mois pour chacune des variables est illustré dans la Table 24

var	IMC	G0	HDL	INS	Quicki	TRYG
1st.quartile_V3M	-19.5%	-22.0%	-10.8%	-61.3%	4.0%	-35.0%
Median_V3M	-16.7%	-9.6%	0.9%	-40.7%	10.9%	-17.0%
3rd.quartile_V3M	-13.1%	-1.0%	12.0%	-19.5%	17.5%	12.5%
1st.quartile_V6M	-27.4%	-25.0%	-6.7%	-70.0%	0.7%	-43.8%
Median_V6M	-23.4%	-15.2%	7.7%	-51.0%	2.1%	-20.6%
3rd.quartile_V6M	-18.7%	-4.7%	20.2%	-34.5%	4.2%	0.0%

Table 24 : mediane, 1er et 3^{ième} quartile des variations à 3 et 6 mois de variables biocliniques

Dans le cadre de cette première analyse du problème, notre objectif était de trouver des modèles de prédiction capables de déterminer la classe des patients en fonction des variables préopératoires les plus importantes et ceci pour chacune des variables d'intérêt. Ainsi, nous avons opté pour l'application de modèles simples et facilement interprétables, de ce fait notre choix s'est fixé sur les arbres décisions qui répondent à nos critères. Pour ce travail nous avons utilisé l'outil de datamining clémentine développé par SPSS.

Pour chacune des variables d'intérêt, nous avons un arbre de décision pour chaque critère de construction de classe et ceci pour la variation à 3 mois et à 6 mois. La Figure 37 reprend les résultats de la prédiction pour chacune des variables classés dans l'ordre décroisant des performances. Ces résultats sont des estimations obtenues par validation croisée (n=10). Nous remarquons que dans la plupart des cas, les meilleurs résultats sont obtenus pour la prédiction des classes extrêmes. Nous avons choisis 90% de précision comme seuil acceptable pour cette première analyse. Pour ce seuil, seul la glycémie et le triglycéride sont retenue et ceci pour la prédiction à 3 mois et à 6 mois des classes extrêmes.



	GLY_3M	TG_3M	TG_6M	GLY_6M	INS_6M	INS_3M	QUICKI_3M	IMC_3M	QUICKI_6M	IMC_6M	HDL_3M	HDL_6M
2C	76.4%	76.5%	81.9%	74.5%	68.4%	63.0%	61.0%	61.0%	53.0%	59.8%	55.6%	62.7%
3C	63.7%	52.3%	65.5%	63.8%	53.7%	51.7%	51.1%	49.0%	37.0%	35.5%	42.8%	42.8%
2C_EXT	91.8%	91.8%	91.2%	90.5%	85.4%	76.4%	72.9%	65.7%	57.0%	54.6%	54.3%	54.3%

Figure 37: Prédiction de la variation des paramètres biocliniques à 3 mois et 6 mois à partir des paramètres préopératoires.

Nous présentons les arbres de décision de la prédiction de variation de la glycémie et du triglycéride respectivement dans la Figure 38 et la Figure 39. Nous remarquons que pour ces variables, la prédiction des classes se fait à partir de la seule valeur de la même variable avant l'intervention chirurgicale.

Les arbres de la Figure 38 sont traduits par les règles suivantes :

- Si la glycémie avant la chirurgie d'un sujet est supérieure à 1.030 alors le sujet appartient à la classe (+, c.à.d. bonne évolution de la glycémie à 3 mois) sinon il appartient à la classe (-, c.à.d. mauvaise évolution de la glycémie à 3 mois)

- Si la glycémie avant la chirurgie d'un sujet est supérieure à 1.070 alors le sujet appartient à la classe (+, c.à.d. bonne évolution de la glycémie à 6 mois) sinon il appartient à la classe (-, c.à.d. mauvaise évolution de la glycémie à 6 mois)

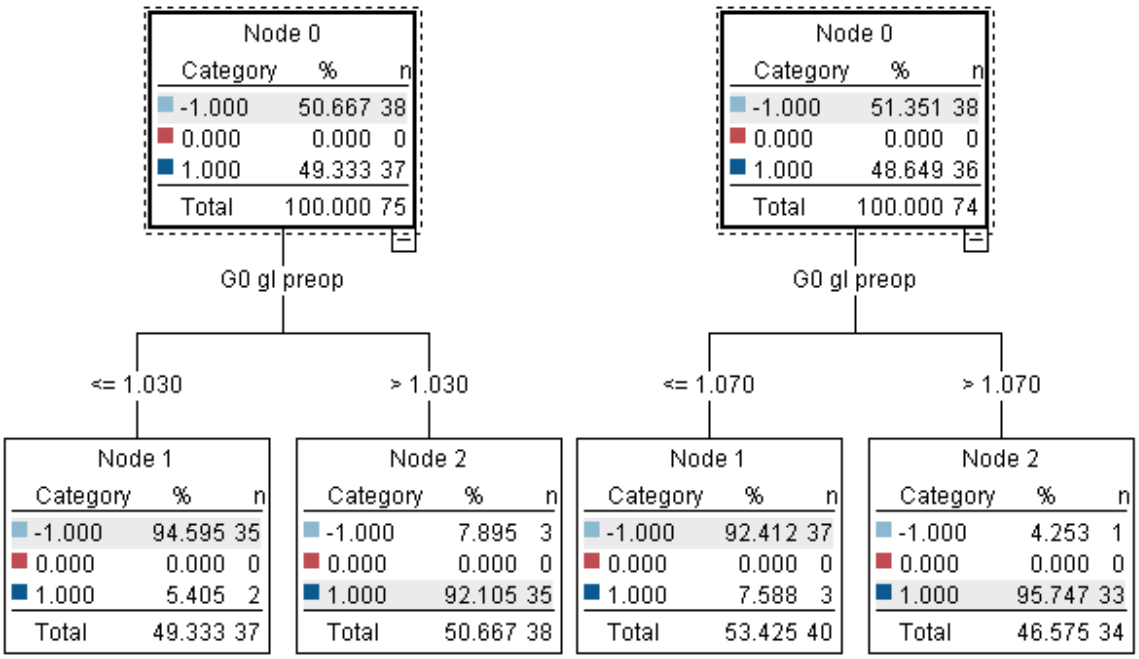


Figure 38 : arbre de décision pour la prédiction de la variation de la glycémie à 3 mois et à 6 mois

Les arbres de la Figure 39 sont traduits par les règles suivantes :

- Si le triglycéride avant la chirurgie d'un sujet est supérieure à 1.400 alors le sujet appartient à la classe (+, c.à.d. bonne évolution de le triglycéride à 3 mois) sinon il appartient à la classe (-, c.à.d. mauvaise évolution de le triglycéride à 3 mois)

- Si le triglycéride avant la chirurgie d'un sujet est supérieure à 1.500 alors le sujet appartient à la classe (+, c.à.d. bonne évolution de le triglycéride à 6 mois) sinon il appartient à la classe (-, c.à.d. mauvaise évolution de le triglycéride à 6 mois)

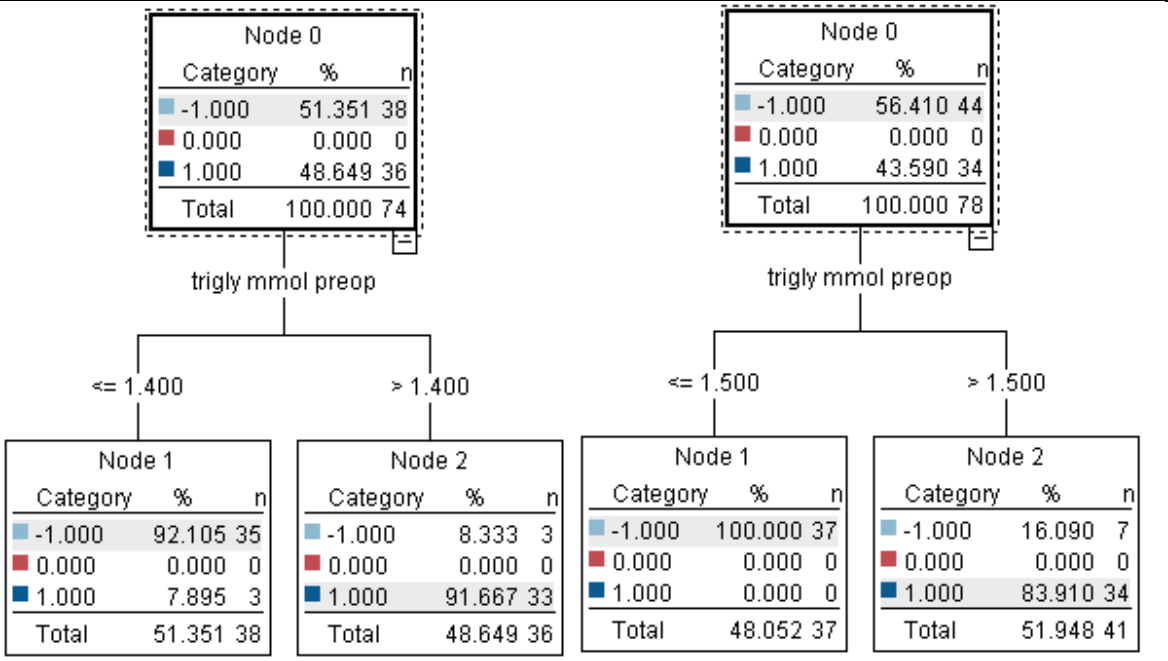


Figure 39 : arbre de décision de la prédiction de la variation du triglycéride à 3 mois et à 6 mois

4.4.3.1 **Analyse des profils d'évolutions des paramètres biocliniques**

Nous avons voulu par la suite combiner les variations de variables biocliniques à 3 mois et à 6 mois pour construire des profils d'évolution et ceci pour le cas des classes extrêmes pour lequel nous avions des résultats meilleurs que les autres critères de création de classes. La Figure 40 reprend le schéma général du flux permettant la création et l'analyse des profils.

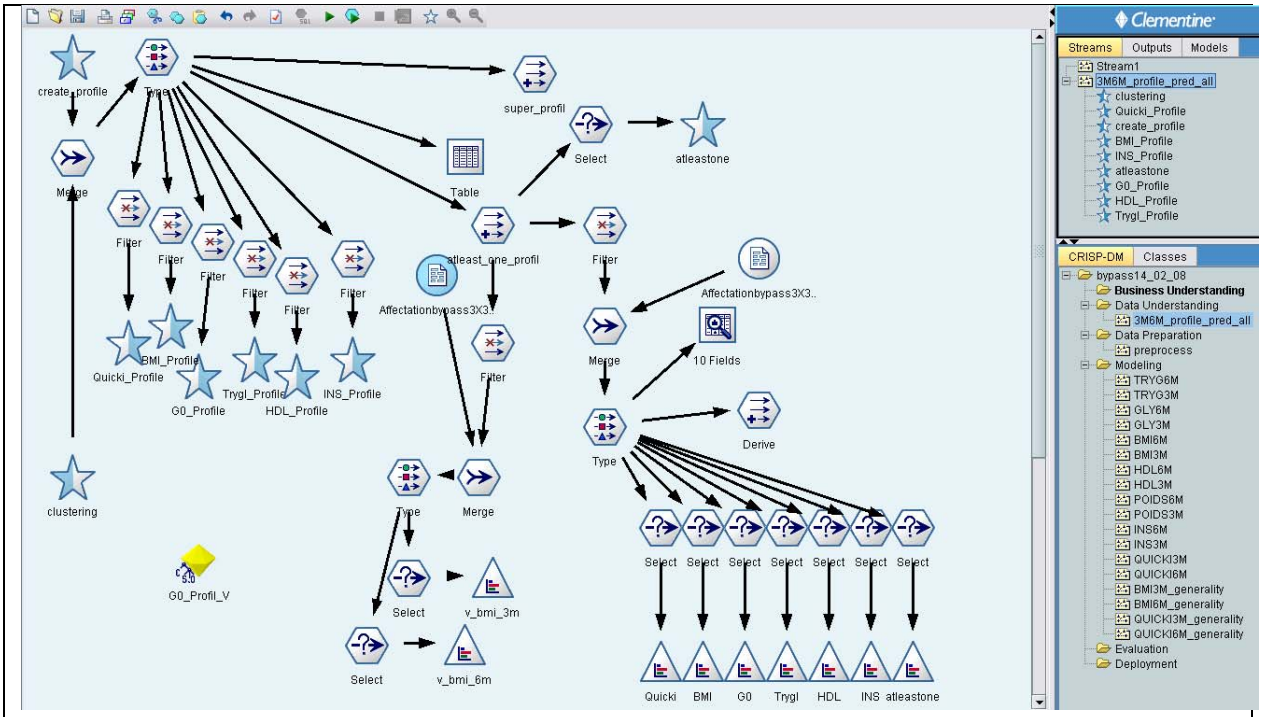


Figure 40: flux de l'analyse des profils de variation des variables biocliniques

Un sujet a un bon profil s'il est dans la classe positive à 3 mois et à 6 mois. Un sujet a un mauvais profil s'il est dans la classe négative à 3 mois et à 6 mois. Une exception est faite la aussi pour le HDL ou l'évolution se fait d'une manière inversée.

Les résultats de la prédiction des profils sont les suivants :

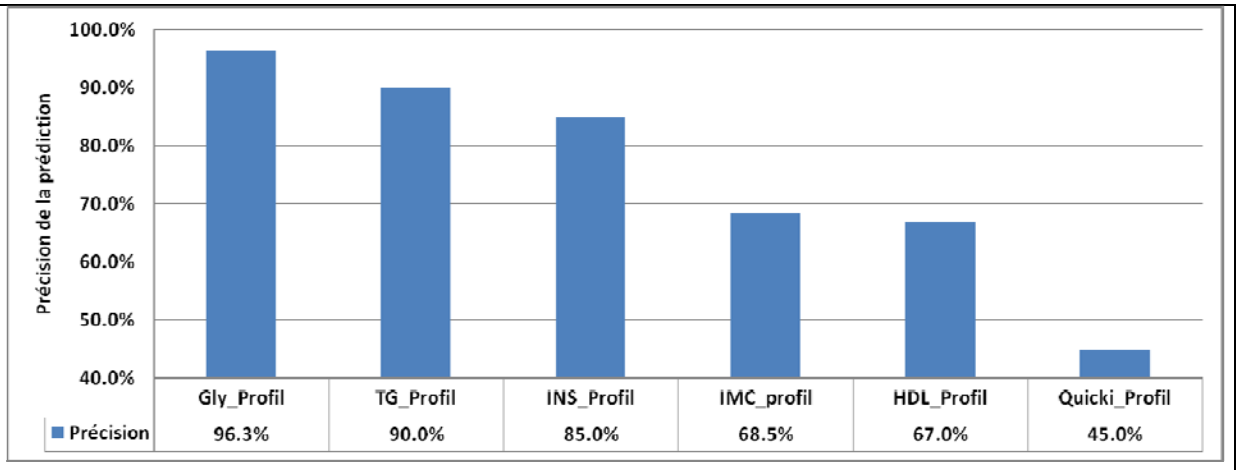


Figure 41 : prédiction du profil de variation des paramètres biocliniques à partir des paramètres préopératoires

Nous constatons là encore que les meilleurs résultats sont obtenus pour la glycémie et le triglycéride.

Enfin, nous avons voulu tester la prédictibilité du profil de patients qui ont une amélioration d'au moins un de leurs paramètres biocliniques. Pour ce modèle, l'estimation de la précision est de 67.6% obtenue avec l'arbre de la Figure 42. Les variables les plus importantes pour la prédiction de ce profil sont : la glycémie, les triglycérides, l'âge de la chirurgie et l'adipondectine.

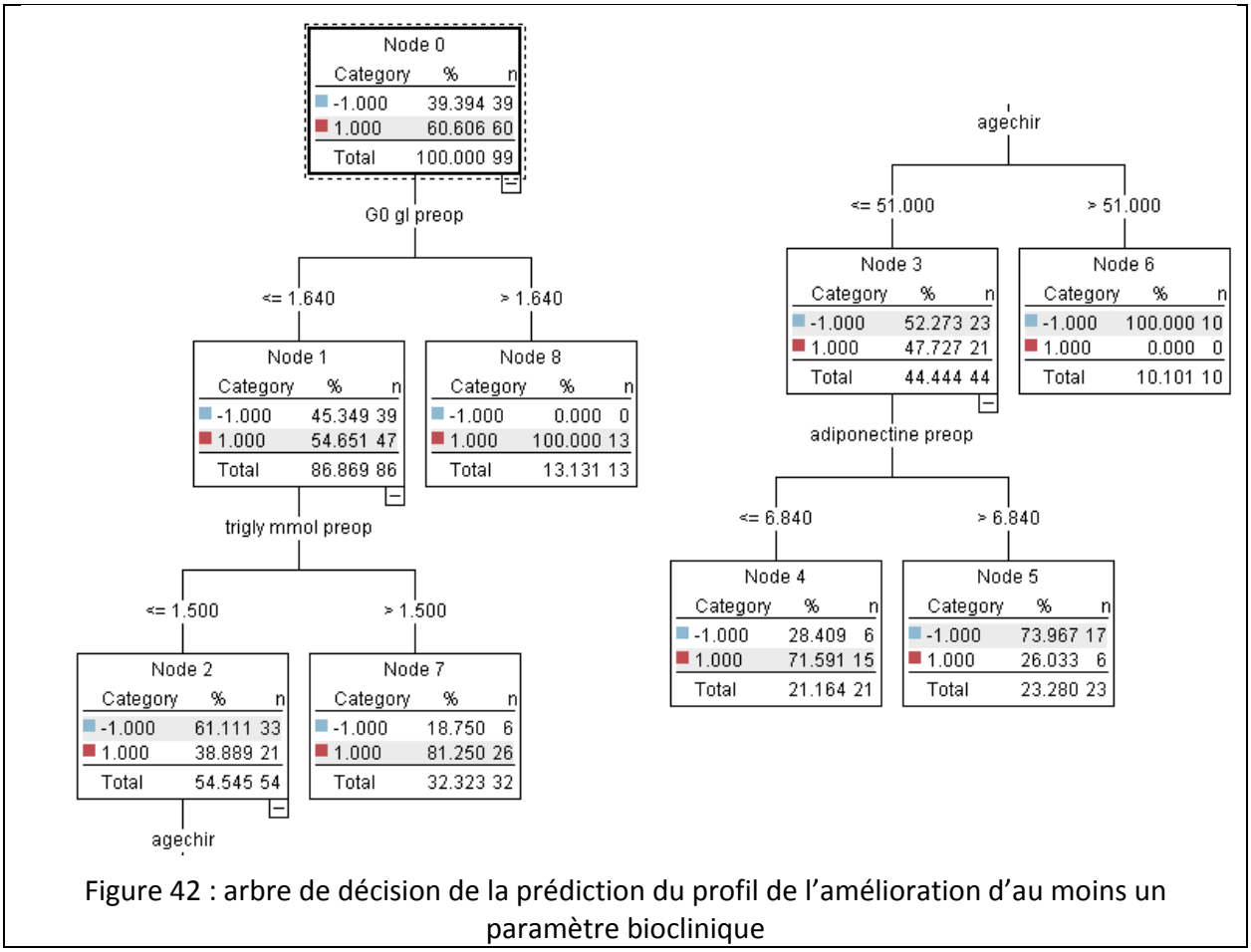


Figure 42 : arbre de décision de la prédiction du profil de l'amélioration d'au moins un paramètre bioclinique

Les règles de décision déduites à partir de cet arbre ci-dessus :

- Si avant la chirurgie, la glycémie d'un sujet est supérieure à 1.640 alors le sujet appartient à la classe (+, c.à.d. bon profil d'évolution)

- Si avant la chirurgie, la glycémie d'un sujet est inférieure ou égale à 1.640 et le triglycéride est supérieure à 1.500 alors le sujet appartient à la classe (+, c.à.d. bon profil d'évolution)

- Si avant la chirurgie, la glycémie d'un sujet est inférieure ou égale à 1.640 et le triglycéride est inférieure ou égale à 1.500 et l'âge est supérieure à 51 alors le sujet appartient à la classe (-, c.à.d. mauvais profil d'évolution)

- Si avant la chirurgie, la glycémie d'un sujet est inférieure ou égale à 1.640 et le triglycéride est inférieure ou égale à 1.500 et l'âge est inférieure ou égale à 51 et l'adiponectine est supérieure à 6.840 alors le sujet appartient à la classe (-, c.à.d. mauvais profil d'évolution)

- Si avant la chirurgie, la glycémie d'un sujet est inférieure ou égale à 1.640 et le triglycéride est inférieure ou égale à 1.500 et l'âge est inférieure ou égale à 51 et l'adiponectine est inférieure ou égale à 6.840 alors le sujet appartient à la classe (-, c.à.d. mauvais profil d'évolution)

Si avant la chirurgie, la glycémie d'un sujet est inférieure ou égale à 1.640 et le triglycéride est inférieure ou égale à 1.500 et l'âge est inférieure ou égale à 51 et l'adiponectine est inférieure ou égale 6.840 alors il appartient à la classe (+, c.à.d. mauvais profil d'évolution)

Profil Glycémie

variables préop	profil(-)	profil(+)	p-value
Glycémie	0.867	1.864	1
Quicki	0.329	0.287	1
Triglycéride	1.33	1.891	0.961

Profil HDL

variables préop	profil(-)	profil(+)	p-value
Leptine preop	66.186	49.089	0.969
CholT	4.622	5.268	0.982
SAApop	38.007	17.74	0.997
HDL	1.546	1.137	0.999

Profil IMC

variables préop	profil(-)	profil(+)	p-value
Age chirurgie	46.625	35.087	1
CRPuspop	0.659	1.115	0.971

Profil Insulénimie

variables préop	profil(-)	profil(+)	p-value
Insuline IRMA	8.986	36.63	1
Quicki	0.34	0.283	1
Adiponectine	7.444	5.207	0.969
HDL	1.405	1.13	0.957

Profil Trycléciride

variables préop	profil(-)	profil(+)	p-value
Triglycéride	0.922	2.588	1
Quicki	0.334	0.301	0.999
Insuline IRMA	12.908	22.322	0.998
ASAT	20.111	28.318	0.996
GGT	33.077	75.227	0.996
ALAT	25.148	45.455	0.995
Adiponectine	8.066	5.492	0.992
HDL	1.492	1.163	0.989
Glycémie	0.982	1.319	0.988
Leptine	64.272	49.882	0.97

Profil Amelioration au moins un paramètre

variables préop	profil(-)	profil(+)	p-value
G0	0.973	1.313	0.999
adiponectine	9.635	6.107	0.999
HDL	1.494	1.226	0.999
trigly	1.223	1.848	0.999
quicki	0.33	0.311	0.996

Table 25 : analyse de variance des variables préopératoires dans les profils biocliniques

Nous avons ensuite réalisé une analyse de variance des variables préopératoires pour chacun des profils et avons retenu celle avec une p-value supérieure à 0.95. Les résultats sont indiqués dans la Table 25.

4.4.3.2 **Discussion**

Nous avons essayé dans cette partie de prédire l'amélioration de l'état de santé des patients suite à un régime et ceci en regardant le profil d'évolution des paramètres biocliniques. Nous avons constaté que l'évolution de certaines variables comme la glycémie et le triglycéride suite à une intervention de type bypass est plus prédictible que l'évolution de variables comme le poids ou l'IMC. Malgré une précision de la prédiction dépassant les 90%, ses modèles de prédiction sont peu informatifs d'un point de vu biologique. La prédiction de la variation de la glycémie se fait à partir de glycémie préopératoire uniquement et il en est de même pour la variation du triglycéride. Pour cette intervention chirurgicale, le suivi de l'évolution du profil de santé des patients à long terme est très important. De ce fait, la prédiction du profil d'évolution temporelle est plus importante qu'une prédiction ponctuelle. L'évolution de la base de données de la chirurgie nous permettra de faire une analyse plus robuste des profils d'évolution à long terme puisque c'est l'amélioration des paramètres biocliniques qui intéresse les praticiens et la connaissance de l'influence d'une intervention chirurgicale sur l'amélioration de la santé des patients obèses serait d'une grande utilité en clinique.

4.5 Conclusion

Dans ce chapitre, nous avons détaillé les analyses que nous avons réalisées au cours des différents projets dans lesquels notre équipe est impliquée. Dans le cadre des deux projets européens, nous avons évalué l'apport des méthodes d'apprentissage supervisé dans la prédiction de la réussite de régimes faibles en calories suivis par des sujets obèses. Pour la construction de ces modèles de prédiction, nous avons eu recours aux données transcriptomiques extraites à partir du tissu adipeux. Dans le projet Nugenob, nous avons observé que le profil transcriptomique prédit faiblement la perte de poids suite à un régime

faible en calorie de 10 semaines. Dans le cadre de ce même projet, nous avons effectué une étude comparative entre la prédiction à partir des données de l'obésité et celles du cancer et nous avons observé de meilleurs résultats prédictifs dans le cadre de l'étude relative au cancer. Dans le projet Diogenes, nous avons observé de bons résultats de prédiction. Ces données nous ont permis de déterminer une liste de prédicateurs que nous avons explorés et analysés. Nous avons une différence significative des résultats de prédiction entre l'étude Nugenob et Diogenes. Nous avons écarté certaines hypothèses qui auraient pu être à l'origine de cette différence mais nous n'avons pas assez d'éléments pour expliquer pour l'instant cette différence.

Dans le cadre de la prédiction de la perte de poids suite à une chirurgie gastrique, nous avons utilisé une base de données bioclinique avec un nombre plus important de sujets. Dans cette étude, en plus de la prédiction de la perte de poids, nous avons essayé de prédire d'autres paramètres biocliniques qui reflètent l'amélioration de l'état de santé des patients. Notre analyse montre que le triglycéride et la glycémie sont prédits avec une meilleure précision que la perte de poids. Nous avons ensuite mis en place des profils d'évolution de ces paramètres à traves le temps et avons défini aussi un paramètre global de l'amélioration que nous avons ensuite essayé de prédire.

Les études montrent une différence quant à la contribution et l'apport des données biocliniques et les données transcriptomiques. Cette hétérogénéité des données peut être exploitée pour l'amélioration de la précision des modèles et aussi pour une meilleure compréhension de la contribution de chaque type de données. Dans ce qui suit, nous présentons deux approches où nous combinons les données cliniques et transcriptomiques.

Chapitre 5

Améliorer la prédiction à partir de la combinaison de données cliniques et transcriptomiques

5.1 Combinaison de données pour l'apprentissage à partir des données biomédicales

On dispose aujourd'hui, dans le domaine de l'apprentissage automatique, d'une multitude de type de classeurs et de méthodes pour les construire. Malgré les nombreux travaux dans le domaine, il est difficile de mettre en évidence la supériorité d'une approche de classification sur une autre ou d'une méthode de prédiction par rapport à une autre (Wolpert 1996; Wolpert 1996)et ce plus particulièrement avec les données de puces à ADN.

Si les méthodes d'intégration de données ont suscité l'intérêt de la communauté d'apprentissage ces dernières années, c'est du fait de l'apparition de données provenant de diverses sources. L'objectif des méthodes d'intégration est la quête de meilleures performances en termes de classification, de prédiction et d'optimisation.

Dans son papier "Multiple classifier combination: Lessons and the next steps" (Ho 2002), Ho a passé en revue les méthodes de combinaison et met l'accent sur l'intérêt de 'bien' utiliser les classeurs et les méthodologies de bases avant de penser à mettre en place une nouvelle approche basée sur la combinaison de modèles. Cela dit, Dietterich (Dietterich 2000) suggère trois raisons pour l'utilisation d'un ensemble de classeurs à la place d'un classeur unique :

- La première est **statistique :** un problème statistique se pose lorsque le volume des données disponibles est trop faible par rapport à la taille de l'espace d'hypothèse. En l'absence d'une quantité suffisante de données, un algorithme

d'apprentissage peut trouver différentes hypothèses qui donnent la même précision sur les données d'apprentissage. En construisant un ensemble de tous ces classeurs, l'algorithme peut "moyenner" le vote de ces classeurs et réduire ainsi le risque de choisir le mauvais classeur.

• La seconde est **computationnelle :** de nombreux classeurs fonctionnent en effectuant des recherches locales qui aboutissent à des optima locaux. Un ensemble construit en exécutant une recherche locale peut fournir une meilleure approximation de la véritable fonction de décision que n'importe lequel des classeurs prit séparément.

• la troisième est **représentationnelle :** dans la plupart des applications liées à l'apprentissage automatique et à la reconnaissance des formes, la fonction de décision ne peut être représentée par aucune des hypothèses possibles. En effectuant une somme pondérée des hypothèses, il est possible d'élargir l'espace de représentation des fonctions.

Nous allons détailler dans ce qui suit les approches de combinaisons et leurs applications dans les différents domaines et nous nous intéresserons en particulier aux approches de combinaison dans le domaine biomédical et en particulier au domaine de la transcriptomique et des modèles de prédiction à partir des puces à ADN.

5.1.1 **Terminologie employée pour la combinaison de données**

Les premières propositions conceptuelles et théoriques de modèles de combinaison furent utilisées au 18^{e} siècle dans les systèmes de vote aux élections. La première étude mathématique des systèmes électoraux par Borda (Borda 1781) et Condorcet (Condorcet 1785) date de la révolution française. Ils ont développé des méthodes de vote capable de mieux réponde aux souhaits des votants. En 1989, Clemen citait déjà 209 travaux reliés à la combinaison de classeurs (Clemen 1989).

Plusieurs propositions de terminologies sont utilisées dans la littérature pour désigner la combinaison. Parmi ces terminologies, on retrouve des appellations telles que les systèmes *multi-classeurs* [multiple classifier systems] (Xu, Krzyzak et al. 1992; Ho, Hull et al. 1994; Ho 2002; Kittler and Alkoot 2003; Melnik, Vardi et al. 2004), les *analyses statistiques multiples* [multiple statistical analysis] (Chuang, Liu et al. 2004; Chuang, Liu et al. 2004; Kuriakose, Chen et al. 2004), le *classement multi-critère* [multi-criterion ranking] (Patil and Taillie 2004), les *systèmes hybrides* [hybrid systems] (Duerr, Haettich et al. 1980; Perrone and Cooper 1993), la *classification d'ensemble* (Xu, Krzyzak et al. 1992; Tumer and Ghosh 1999; Ho 2002; Kittler and Alkoot 2003), la *combinaison d'évidence* (Belkin, Kantor et al. 1995; Chuang, Liu et al. 2004; Chuang, Liu et al. 2004), la *fusion de données/d'informations* (Dasarathy 2000; Ibraev, Kantor et al. 2001; Ibraev, Ng et al. 2002; Hsu and Taksa 2005), l'*agrégation de rang* (Dwork, Kumar et al. 2001; Fagin, Kumar et al. 2003), le scoring de consensus (Gohlke and Klebe 2001; Clark, Strizhev et al. 2002; Yang, Chen et al. 2005).

Malgré une littérature abondante et une multitude de travaux qui ont abordé le sujet de différentes manières, le concepteur d'un système de combinaison de classeurs est toujours confronté à un certain nombre de choix auxquels la communauté de l'apprentissage et de la classification n'a pas encore donné de réponses claires. D'une part, le problème est complexe à modéliser et très peu de travaux théoriques existent sur le sujet. D'autre part, on retrouve des études empiriques qui traitent des différents problèmes mais les résultats obtenus restent étroitement liés aux problèmes étudiés et par conséquent il est difficile de généraliser ces approches et de les appliquer directement dans d'autres domaines.

La multiplication des travaux sur la combinaison a entraîné la mise au point de nombreux schémas traitant les données de manières différentes (Kuncheva 2004).

5.1.2 **Les différentes stratégies de combinaisons**

Plusieurs stratégies de combinaisons sont proposées dans la littérature afin de regrouper les familles de méthodes de combinaison et nous retenons les trois stratégies suivantes : une stratégie selon *le niveau* de combinaison, une stratégie selon la *structure* de la combinaison et une stratégie selon le type de sortie des algorithmes.

5.1.2.1 **Les niveaux de combinaisons**

Dans son livre « Multiple classifier system », Ludmilla Kuncheva regroupe les modèles de combinaison en 4 grands niveaux de combinaison à savoir le niveau de combinaison de données, la combinaison d'attributs, la combinaison de classeurs et la combinaison de modèles de combinaison de classeurs.

- **La combinaison de données :** Les chercheurs ont très souvent concentré leurs travaux sur l'amélioration d'un classeur unique principalement en raison de leur manque de ressources suffisantes pour développer simultanément plusieurs classeurs. Une méthode simple pour générer plusieurs classeurs dans ce cas est de lancer plusieurs sessions avec un même classeur et différents sous-ensembles de données. La première approche basée sur cette idée a été proposée par Breiman (Breiman 1996) est connu sous le nom de "Bagging". Cette méthode construit les ensembles établis en effectuant des tirages avec remplacements de l'ensemble original. Chaque ensemble donnant lieu à une classification un peu différente. La technique utilisée pour générer les différents ensembles d'apprentissages est également connue sous le nom de bootstrap et vise à réduire l'erreur des estimateurs statistiques. Dans la pratique, le bagging a montré des bons résultats. Toutefois, les gains de performance sont généralement faibles lorsque le bagging est appliqué à des classeurs faibles. Dans ces cas, une autre technique est à déployer : le boosting.
- **La combinaison d'attributs :** il s'agit d'utiliser plusieurs sous ensemble d'attributs pour les classeurs
- **La combinaison de classeurs :** plusieurs classeurs peuvent être utilisés comme classeur basique pour la combinaison. Le modèle de classeurs est choisi selon l'interopérabilité de leur décision, la facilité d'implémentation et l'adaptabilité à d'autre problèmes (Webb 2002).
- **La combinaison de décision de classeurs :** il s'agit dans ce cas de combiner la décision des classeurs.

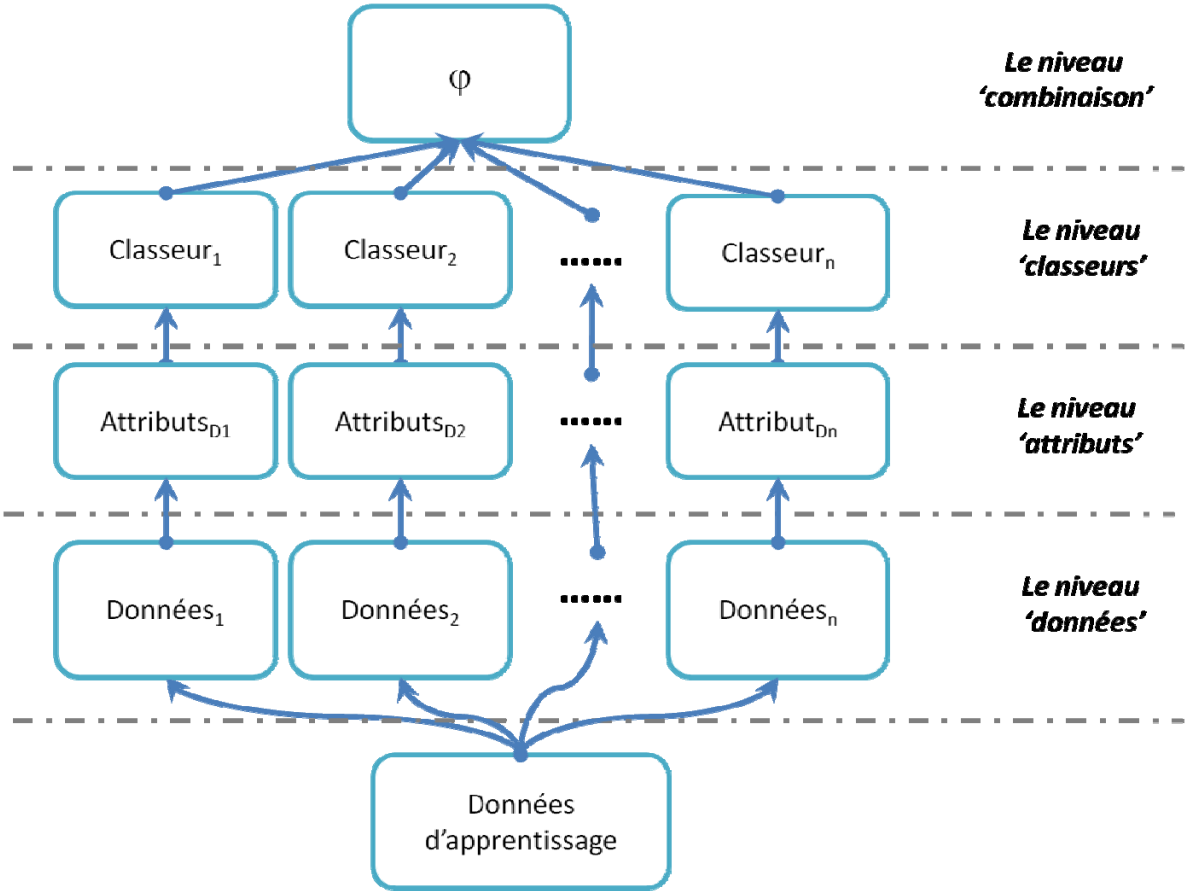


Figure 43: Niveaux de construction de modèle de combinaison d'après (Ludmila I. Kuncheva 2004)

5.1.2.2 **Les structures de la combinaison**

Un autre regroupement possible des classeurs est le regroupement structurel :

- **La combinaison séquentielle :** La combinaison séquentielle (série) est organisée en niveaux successifs de décision permettant de réduire progressivement le nombre de classes possibles. Dans chaque niveau, il existe un seul classeur qui prend en compte la réponse fournie par le classeur placé en amont pour traiter les rejets ou confirmer la décision obtenue sur la forme qui lui est présentée. Une telle approche peut être vue comme un filtrage des décisions. Notons tous de même que cette approche est sensible

à l'ordre. En effet, la disposition des classeurs influe les performances du modèle de combinaison.

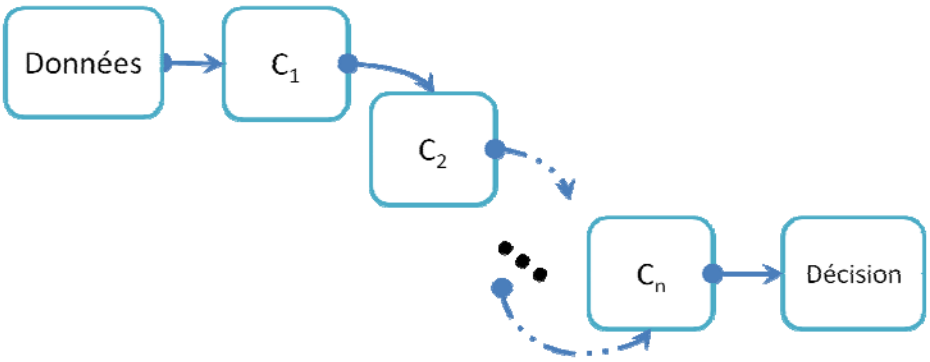


Figure 44: Combinaison série de classeurs

- **La combinaison parallèle:** Dans cette approche, les différents classeurs opèrent indépendamment les uns des autres puis fusionnent leurs sorties respectives. Cette fusion peut avoir lieu comme elle peut avoir lieu avec une pondération favorisant un algorithme par rapport à un autre selon l'aptitude de ce dernier.

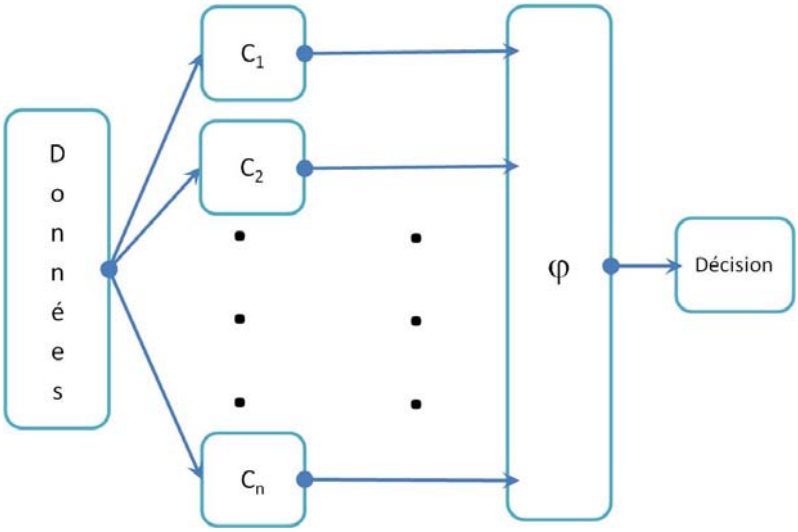


Figure 45: Combinaison parallèle de classeurs

- **La combinaison hybride :** La combinaison hybride intègre à la fois la combinaison séquentielle et la combinaison parallèle.

- **La combinaison Hiérarchique :** La combinaison hiérarchique consiste à combiner les classeurs en cascade, de sorte que la sortie des classeurs de base soit utilisée comme entrée dans le classeur du nœud parent.

La combinaison parallèle est la structure de combinaison la plus utilisée du fait de sa simplicité et sa stabilité par rapport au classeur de base.

5.1.2.3 **Les types de combinaisons**

Un autre moyen de regrouper la famille des classeurs peut se baser sur le type de la sortie d'un classeur de base. Dans (Xu, Krzyzak et al. 1992), trois différents types de sortie sont définis et ainsi trois types de combinaison sont proposés :

- **Le type abstrait (classe) :** Dans ce cas de figure, chaque classeur produit en sortie une étiquette qui reflète la classe d'appartenance de l'instance sans pour autant fournir une information sur la confiance que le modèle attache a cette prédiction. Ce type est le plus général puisque chaque classeur, peu importe sa nature, fournit une classe pour les instances prédites

- **Le type rang :** Dans ce cas, les classeurs fournissent un classement sur les classes dans l'ordre de préférence des classes selon la confiance qu'accorde le classeur à chaque classe (Tubbs and Alltop 1991; Ho, Hull et al. 1994).

- **Le type mesure :** Dans ce cas, le classeur fournit un degré de confiance en chaque classe. C'est une estimation (ou une mesure qui reflète une estimation) de la probabilité qu'une instance soit étiquetée par le label de la classe N**.**

Chaque type de sortie (classe, rang ou mesure) correspond à un niveau d'information que nous fournit le classeur. La sortie de type classe est la plus simple mais la plus limitée aussi. La sortie de type rang donne un ordre de préférence des propositions fournies par le classeur. La sortie de type mesure est la plus riche en information.

Un quatrième type introduit par (Kuncheva 2004) est le **type oracle.** Dans ce cas le classeur donne une réponse (bonne ou mauvaise prédiction) sans avoir des connaissances sur l'appartenance de l'instance à une classe.

5.1.2.4 **Les règles de combinaison**

Il existe deux catégories de règles de combinaison : la combinaison non-paramétrique (règle fixe) et la combinaison paramétrique (pondérée). Soit ϕ la fonction de décision résultante de la combinaison des classeurs $C_1,...,C_n$. La fonction peut s'écrire $\varphi = F(C_1,...,C_n)$ où F est la règle de combinaison.

- **La combinaison non-paramétrique :** utilise les informations (sorties de classeurs) de manière égale. Elles sont faciles à implémenter. Cette famille regroupe les fonctions arithmétiques courantes comme la moyenne ($\varphi = \frac{1}{n}\sum_{i=1}^{n} C_i$), le maximum, le minimum, la médiane, la somme et le produit ($\varphi = \frac{1}{n}\prod_{i=1}^{n} C_i$). D'après (Kittler, Hatef et al. 1998), la somme est plus résistante aux erreurs de classeur de base que les autres règles fixes alors que le produit et le minimum sont les plus sensibles aux erreurs.
- **La combinaison paramétrique:** utilise des paramètres supplémentaires qui sont calculés pendant la phase d'apprentissage. Ces paramètres sont en général des poids (coefficients) attribués à chacun des classeurs. Ces poids reflètent en général l'influence de chacun des classeurs de base dans la combinaison. Parmi les méthodes de combinaison paramétriques, les approches de combinaison linéaire (Figure 46) sont les plus utilisées (Kittler, Hatef et al. 1998; Tumer and Ghosh 1999; Tax, van Breukelen et al. 2000; Kuncheva 2002). La fonction de décision φ se traduit par: $\varphi = \sum_{i=1}^{n} \alpha_i . C_i$

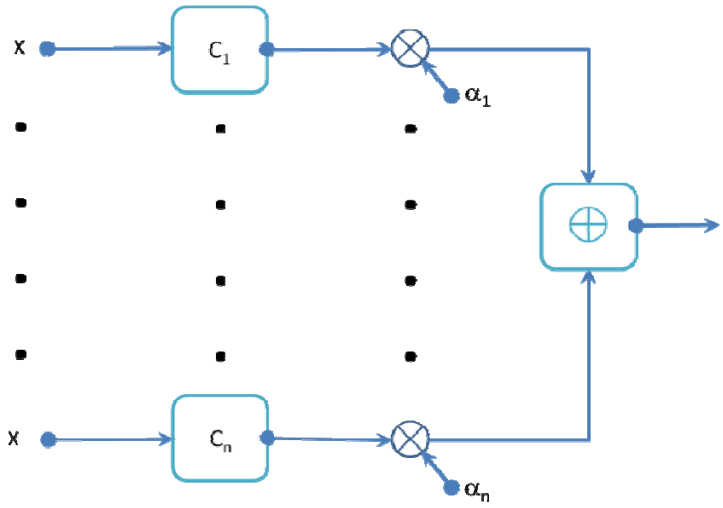


Figure 46 : Combinaison linéaire de n classeurs

La combinaison paramétrique est connue pour être plus pertinente que la combinaison non-paramétrique (Duin and Tax 2000; Roli, Raudys et al. 2002)

Nous avons passé brièvement en revue les différentes stratégies de combinaison présentées dans la littérature, le bilan est favorable pour l'utilisation de la structure de combinaison parallèle avec des règles paramétriques. Dans ce qui suit, nous reprenons les travaux réalisés dans le domaine biomédical avec une attention particulière dirigée vers la combinaison intégrant des données *'omiques'*. Nous nous intéresserons en particulier aux méthodes à noyaux, les machines à vecteur de support, qui sont très répandues dans ce domaine du fait de la bonne performance de ces méthodes.

5.1.3 Classification à partir de la combinaison de données dans le domaine biomédical avec les machines à vecteurs de support

Le nombre de bases de données biomédicales suit une évolution exponentielle. Par exemple, la base de données du journal « Nucleic Acids Research » est passée de 96 bases en 2001 (Baxevanis 2001) à plus de 800 bases en 2007 (Galperin 2007). Cette évolution et la disponibilité des ressources et des données patients de diverses natures (génétique, transcriptomique, protéomique, clinique, environnementale, ...) ont accéléré la recherche médicale et le développement des méthodes informatiques à des fins diagnostiques et

pronostiques en particulier dans le domaine de l'oncologie. Si plusieurs approches utilisent soi les données cliniques (Hadzikadic, Hakenewerth et al. 1996; Signorini DF), soi les données biopuces (Golub 1999; Alizadeh, Eisen et al. 2000; Shipp, Ross et al. 2002) pour faire de la prédiction des résultats médicaux, la combinaison de diverses informations pourrait augmenter les performances des modèles obtenus. Plusieurs travaux dans cette optique ont montré l'intérêt de la combinaison.

Un nombre considérable de travaux adresse le problème de l'intégration automatique des données génomiques en considérant les interactions entre les différentes sources de données.

Afin d'appliquer un SVM à une base de données hétérogène, les noyaux doivent être définis pour chaque type de données qui sont ensuite combinées. Par exemple, Pavlidis et al. (Pavlidis, Weston et al. 2001) utilisent cette approche pour combiner les données d'expression géniques et les profils phylogénétiques d'une manière non pondérée. Il propose trois formalismes d'intégration pour prédire la classe fonctionnelle des gènes : *intégration initiale*, *intégration intermédiaire* et *intégration tardive*. Dans l'intégration « initiale », les deux ensembles de données sont concaténés pour former une seule entrée. Dans l'intégration « intermédiaire », les matrices de Gram sont calculées séparément pour chaque ensemble de données et ensuite additionnées. Dans l'intégration « tardive », un SVM est formé à partir de chaque type de données, et les valeurs discriminantes résultantes sont additionnées.

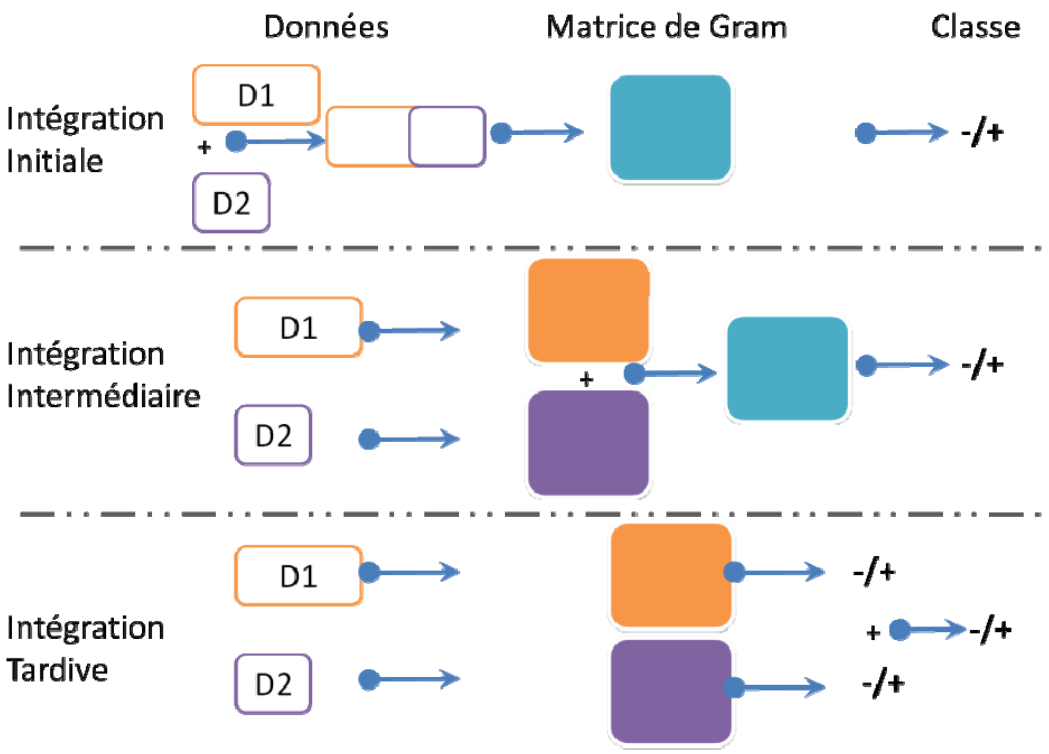


Figure 47: Méthodes d'intégration de données avec les SVM. D'après (Pavlidis, Weston et al. 2001)

Pavlidis démontre que les SVMs peuvent apprendre à partir de données hétérogènes. Avec une fonction noyau appropriée, les machines à vecteur de support peuvent apprendre à partir d'une combinaison de deux sources de données différentes. Pavlidis constate que dans la plupart des cas, les SVM appris à partir de la combinaison fournissent des résultats équivalents ou meilleurs que ceux obtenus avec des modèles appris à partir des données prises séparément. Il prouve également que la méthode intermédiaire d'intégration fournit de résultats plus intéressants que les autres techniques de combinaison qu'il a analysé.

Une autre approche de combinaison de noyaux a été proposée par Lanckriet et al. (Lanckriet 2004). Dans ce travail, les auteurs ont proposé une méthode de prédiction paramétrique basée sur la résolution d'un problème d'optimisation quadratique avec des contraintes quadratiques par l'intermédiaire de la programmation semi-définie. Cette approche

permet de définir les coefficients optimaux de la pondération. Ils combinent des données biopuces, les séquences des protéines et les interactions protéine-protéine. Cette méthode applique les machines à vecteur de support pour prédire les fonctions des protéines chez la levure. Les résultats obtenus ont montré que la combinaison de données conduit à une amélioration de la précision de prédiction par rapport aux méthodes usuelles appliquées sur ce type de données. Cependant, aucune implémentation libre de cette approche n'est disponible à ce jour.

Une somme non pondérée des noyaux a également été utilisée avec succès dans la prédiction des interactions protéine-protéine (Ben-Hur and Noble 2005). Les auteurs combinent les séquences des protéines, les annotations Gene Ontology, les propriétés du réseau d'affinité et les interactions de séquences homologues dans d'autres espèces. Ils utilisent leur méthode pour prédire les interactions physiques dans la levure en utilisant les données de la base BIND. À un taux de faux positifs de 1% le classeur récupère près de 80% de l'ensemble d'interactions fiables.

Dans le cadre de la combinaison de deux sources, Lewis (Lewis, Jebara et al. 2006) réalise une étude empirique de la prédiction de l'annotation fonctionnelle des gènes à partir de la combinaison de la structure des protéines et de la structure des données. Il conclut que dans le cas de la présence de deux sources de données, une combinaison non pondérée par les SVM produit une précision de prédiction comparable aux approches paramétriques plus sophistiquées. Dans ce cas simple, il estime que la combinaison par la moyenne serait suffisante. D'autre part, il constate que, pour des données bruitées l'approche pondérée devient plus intéressante que l'approche simple. L'auteur ne propose pas l'implémentation de son approche à la communauté des chercheurs pour expérimenter ses approches.

Encore une fois, le choix entre l'approche pondérée et non pondérée est très délicat même dans le cas le plus simple de la combinaison de deux sources. D'autres analyses empiriques sont nécessaires pour avoir une vision plus claire du meilleur choix. Il serait aussi intéressant de disposer d'un outil permettant d'observer les performances des différents modèles de combinaison afin de convenir du meilleur choix.

C'est pour combler ce vide dans le domaine de la prédiction à partir de la combinaison de données que nous présentons notre approche 2KC-SVM (2 kernel-combination SVM) qui fournit un outil pour analyser les méthodes de combinaison à deux sources.

5.1.4 **Notre contribution à la combinaison: 2KC-SVM**

La combinaison de noyaux est une forme de combinaison paramétrique qui repose sur une composition des différentes mesures de similarité recueillies à partir de différentes sources de données. On construit en un premier temps les matrices de Gram pour chaque ensemble de données, l'objectif ensuite est de construire une mesure de similarité 'optimale' issue de cette combinaison linéaire. Etant donnée un ensemble de noyaux $\kappa = \{K_1, K_2, ..., K_m\}$, on construit une combinaison linéaire K : $K = \sum_{i=1}^{m} \mu_i K_i, K$ définie positive (9)

Avec $\mu_i \geq 0, \forall i \in [1, m]$, cette condition assure que le noyau résultat *K* satisfait la condition de Mercer.

Le problème s'écrit :

$$\min_{\alpha,\mu} \frac{1}{2} \sum_{i,j=1}^{m} y_i y_j \alpha_i \alpha_j K(\mathbf{x}_i \cdot \mathbf{x}_j) - \sum_{i=1}^{m} \alpha_i$$

$$\text{sous containte} \begin{cases} \sum_{i=1}^{m} y_i \alpha_i = 0, \\ 0 \leq \alpha_i \leq C, \; i \in [1..m] \\ K = \sum_{l=1}^{n} \mu_l K_l, \; K \succ 0 \end{cases} \quad (10)$$

Ce problème s'écrit encore sous la forme suivante :

$$\max_{\alpha,t} 2\alpha^T e - \tau \alpha^T \alpha - ct$$

$$\text{sous containte} \begin{cases} t \geq b\alpha^T G(Ki)\alpha, i \in [1, m] \\ \alpha^T y = 0 \\ 0 \leq \alpha \leq C \end{cases} \quad (11)$$

Ce problème est un problème de programmation quadratique avec des contraintes quadratiques. Plus de détails peuvent être trouvés dans (Lanckriet 2004).

Dans le cas particulier où il existe uniquement deux source de données, soit clinique et expression, deux noyaux K_c et K_e (Figure 48), le noyau résultant s'écrit alors : $K = \mu_c * K_c + \mu_e * K_e$ Cette équation peut se réécrire sous la forme suivante : $K = (1-\mu) * K_c + \mu * K_e$. En fixant la valeur du paramètre $\mu \in [0..1]$, on se ramène à la résolution du SVM standard.

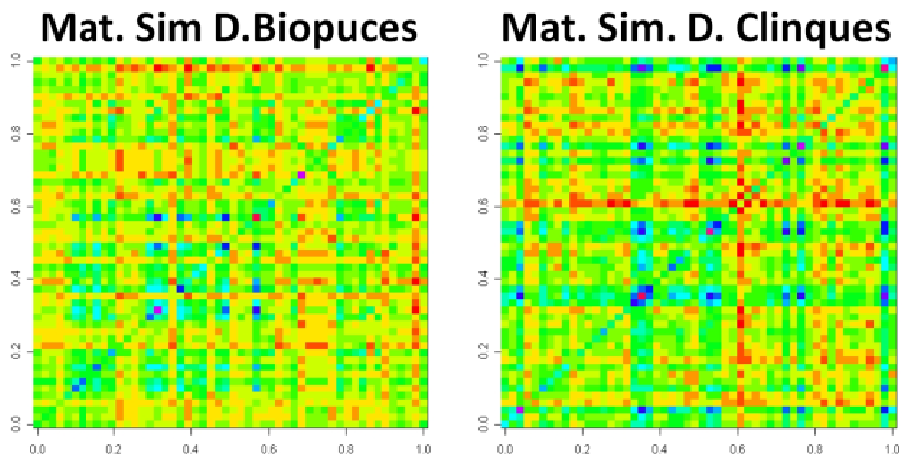


Figure 48 : exemple de matrice de Gram calculée à partir des données biopuce et cliniques de l'étude VLCD

Le parcours des valeurs possible de $\mu \in [0..1]$ permet d'avoir une courbe paramétrique en μ des performances du modèle de combinaison et permet ainsi de comprendre l'impact de la pondération dans la combinaison et la contribution de chacune des sources de données. L'aboutissement au résultat se fait en deux phases : une phase d'apprentissage qui permet de définir une règle de séparation capable de classer et la deuxième c'est laquelle ? Le principe de l'approche est décrit dans ce qui suit :

Algorithme 1 pseudo code de l'algorithme 2KC

D.CLI=[x1…xn] (Premier jeu de données (clinique))

p.DC = [Kernel, C] (Paramètres du noyau pour le premier jeu de données)

D.EXP = [y1 … yn] (Second jeu de données (expression))

p.DE= [Kernel, C] (Paramètres du noyau pour le second jeu de données)

Classe = [l1 … ln] (vecteur avec la classe d'appartenance de chaque entrée)

Pour chaque valeur de μ dans [0..1] par un pas de α=0.05

Diviser (D.CLI, D.CLI_app, D.CLI_test)

Diviser (D.EXP, D.EXP_app, D.EXP_test)

Kern_CLI←Construire_Noyau(D.CLI_app,pDC)

Kern_EXP← Construire_Noyau (D.EXP_app,pDE)

normaliser_Noyau (Kern_CLI)

normaliser_Noyau (Kern_EXP)

KC←Combiner_noyau(Kern_CLI,Kern_EXP,μ)

Apprendre_SVM(KC, Classe,μ)

estimer_SVM(D.CLI_test, D.EXP_test,μ)

Fin boucle

Afficher (estimer_SVM)

L'algorithme de combinaison décrit ci-dessus a été implémenté en utilisant le langage de programmation R (Team 2003). Le code et la procédure d'installation du package sont disponibles à l'adresse suivante : http://temanni.ramzi.free.fr/2KCSVM/

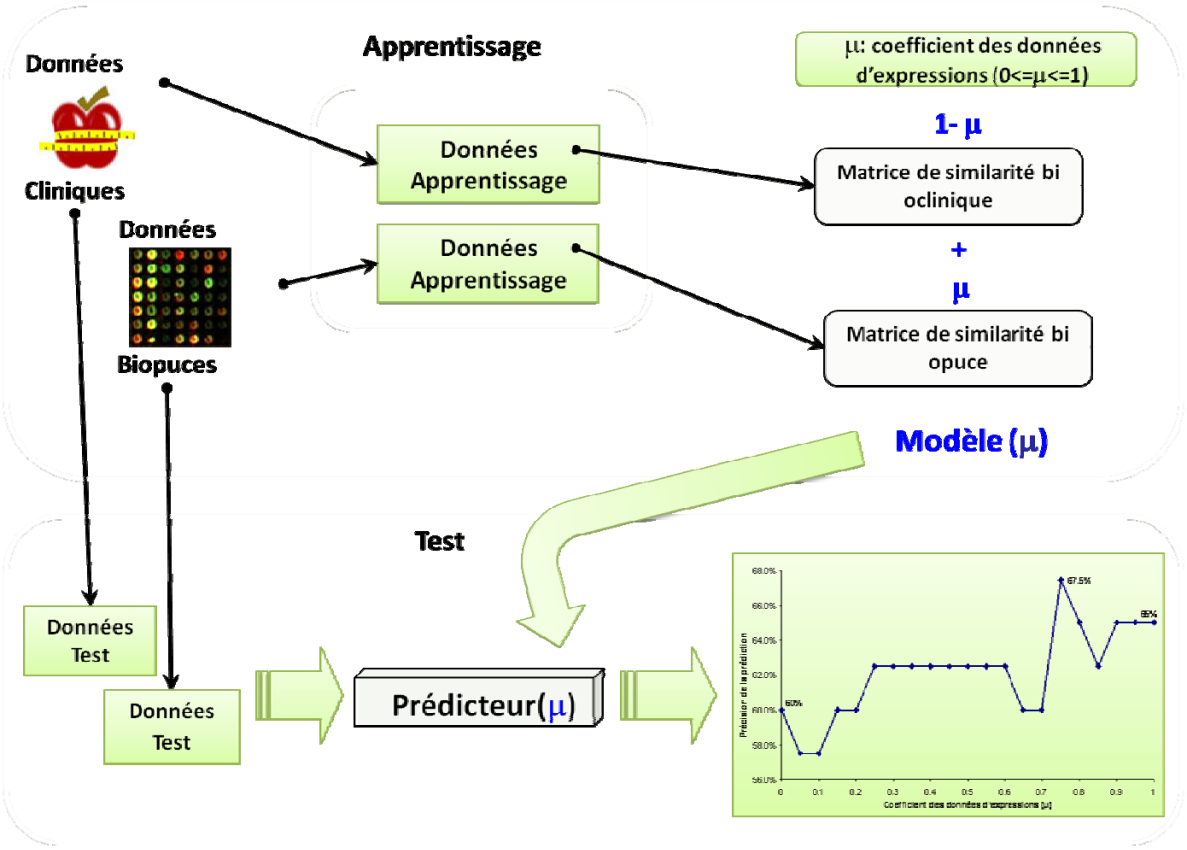


Figure 49 : principe de l'approche 2KC-SVM

5.1.5 **Résultats**

Dans ce qui suit, nous présentons les résultats obtenus par la méthode 2KC-SVM sur les données présentées dans le chapitre 2.5

5.1.5.1 **Données cancer**

Nous présentons ci-après, les résultats obtenus en utilisant notre méthode de combinaison. Nous traçons pour chacun des deux jeux de données utilisés la courbe représentant la variation de la précision de la prédiction selon le critère d'estimation *leave-one-out* en fonction du paramètre μ. Nous avons évalué différents noyaux (linéaire, polynomial de degré 2, polynomial de degré 3 et radial) et différentes valeurs du paramètre de régularisation C

pour chaque source de données. Ensuite, nous avons essayé plusieurs combinaisons de fonctions noyaux (lin-lin, lin-poly2, …, radial-radial) et nous donnons dans ce qui suit seulement la meilleure combinaison pour chaque ensemble de données.

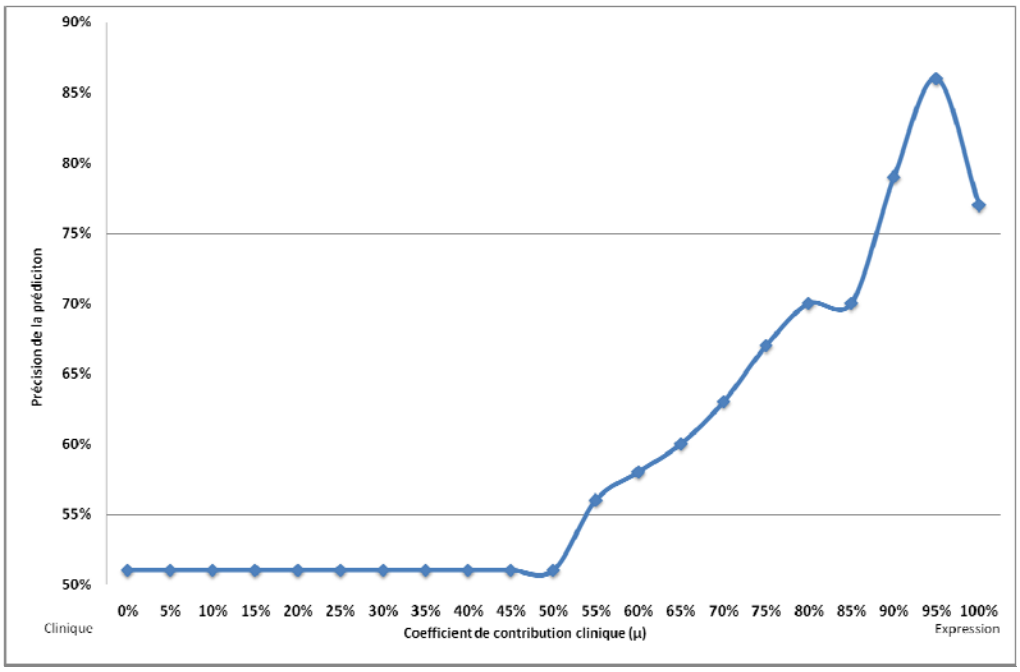


Figure 50 : Variation de la précision de la prédiction (survie après 5 ans, données Harvard)

La Figure 50 représente la courbe de la variation de la précision de la prédiction de la survie après 5 ans pour les données de l'université de Harvard en fonction du paramètre µ. Les résultats obtenus à partir des données cliniques seules (µ=0) est de 51,16%. Avec les données d'expression (µ=1) nous obtenons une précision de 76,74%. Nous avons une valeur constante de la précision pour $\mu \in [0,0.5]$, puis un accroissement pour $\mu \in [0.5,0.95]$. La précision de prédiction optimale est de 86,05%, elle est obtenue pour µ=0,95. Les résultats que nous obtenons avec la combinaison sont améliorés de 10% par comparaison aux résultats obtenus à partir des données d'expressions et de 25% comparé aux résultats obtenus à partir des données cliniques.

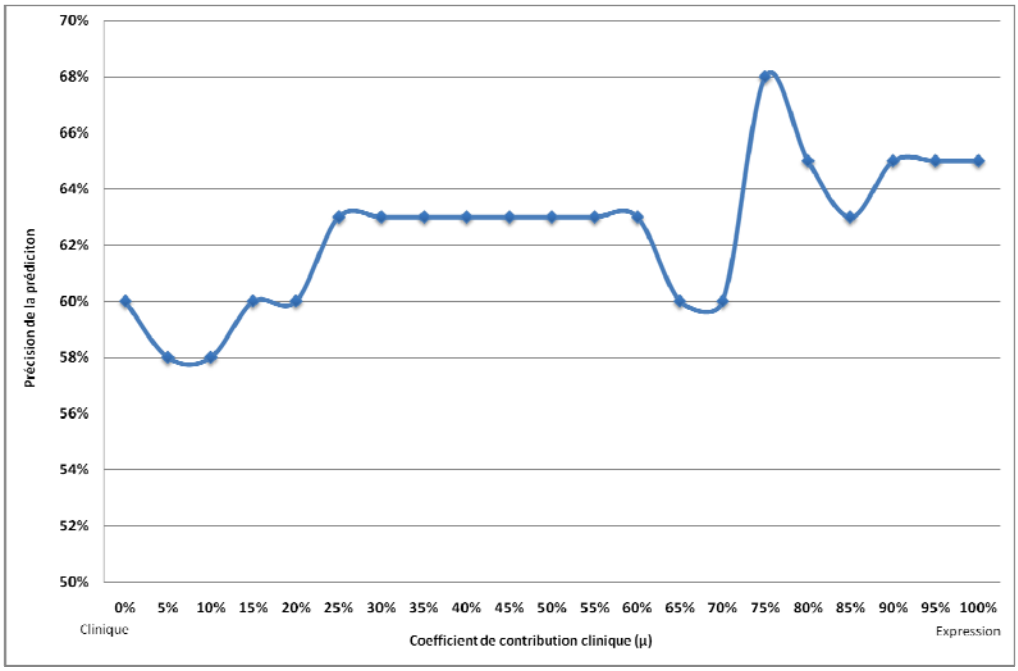


Figure 51 : Variation de la précision de la prédiction (survie après 5 ans, données Massachusetts)

Comme avec le premier jeu de données, la Figure 51 exprime la précision de la prédiction de la survie après cinq ans avec les données de l'université du Massachusetts. La précision obtenue en utilisant uniquement les données cliniques (µ=0) est d'environ 60%, avec les données d'expression (µ=1) la précision est d'environ 65%. Avec les données de la tumeur du cerveau, quelques combinaisons (µ=0.05, µ=0.1) ont des performances inférieures à celles obtenues à partir données cliniques seules, alors que pour les autres valeurs les résultats sont meilleurs. Le meilleur résultat est de 67,5%, ce résultat a été obtenu pour µ=0.75. La combinaison optimale améliore le pouvoir prédictif de 2,5% comparé aux résultats obtenus à partir des données d'expression et d'environ de 7,5% comparé aux résultats obtenus à partir des données cliniques.

La Figure 51 donne la variation de la précision de la prédiction de la survie après 5 ans selon le paramètre coefficient d'expression génique (μ). Nous obtenons le meilleur résultat pour une valeur de 0,7 pour le coefficient clinique et une valeur de 0,3 pour le coefficient

d'expression. La meilleure performance obtenue est de 79% soit 7% de plus que ce que nous obtenons avec les données d'expressions et 28% de plus par rapport aux données cliniques.

5.1.5.2 **Données obésité**

Pour la base VLCD, la Fig. 52 illustre la variation de la précision de la prédiction de la perte de poids pour un seuil de variation de l'indice de masse corporelle de 7%. Ces résultats sont obtenus en combinant les données cliniques avec l'intégralité des données d'expression génique en utilisant un noyau linéaire pour chaque ensemble de données et le paramètre C=1. La précision obtenue en utilisant uniquement les données cliniques (µ=0) est d'environ 64%, avec les données d'expression (µ=1) la précision est d'environ 53%. La précision optimale est obtenue pour une valeur de µ comprise entre 0.7 et 0.9, cette précision est d'environ 72%.

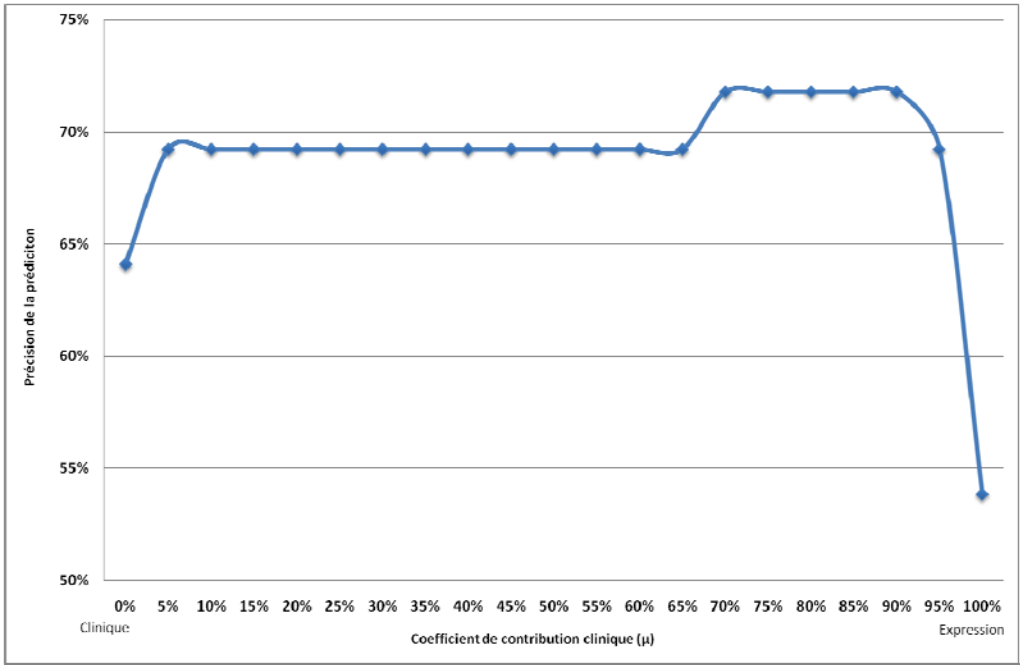


Fig. 52 : Variation de la précision de la prédiction (Perte de poids suite à un régime faible en calories) en combinant les données clinique avec les données d'expression génique

5.1.5.2.1 Sélection de gènes par expérimentation croisée témoin versus obèse

Partant de l'hypothèse que les gènes susceptibles d'intervenir dans la prédiction de la perte de poids chez les obèses seraient des gènes qui s'expriment d'une manière différente chez les sujets sains, nous nous sommes intéressés à l'étude des gènes qui différencient les obèses des personnes de poids normal. Nous avons réalisé une analyse afin de déterminer un profil d'expression qui différentie les 39 sujets obèses d'un pool de 10 sujets témoins normopondéraux. Le résumé des paramètres des patients est indiqué dans la Table 26.

Table 26: Paramètres cliniques et biologiques des sujets obèses et des témoins

	Sujet Obese(39)	Sujet Témoin(10)
Phenotype		
Age (years)	40.36 ± 8.02	35.2 ± 7.69
BMI (kg/m²)	40.58 ± 1.58**	23.67 ± 0.48
homéostasie glucidique		
Glucose (mmol/l)	5.31 ± 0.15	4.82 ± 0.45
Insuline (µU/ml)	12.53 ± 1.38	7.20 ± 1.56
QUICKI	0.33 ± 0.01	0.36 ± 0.02
Diabète type 2		
Glycemie >7 mmol/l or treatment	6 (11%)	0
homéostasie Lipidique		
Cholesterol (mmol/l)	5.42 ± 0.17*	4.36 ± 0.47
HDL cholesterol (mmol/l)	1.28 ± 0.08	1.43 ± 0.1
Triglycerides (mmol/l)	1.40 ± 0.11**	0.45 ± 0.05
Adipokines		
Leptine (ng/ml)	51.84 ± 3.86	11.24 ± 0.35
Adiponectine (µg/ml)	4.54 ± 0.53	—
Facteurs de risque cardiovasculaire		

HDL <1.03 mmol/l (M), <1.29 mmol/l (F)	8 (32%)	1 (10%)
Hypertension ≥130/85 mmHg	8 (32%)	0
Glucose ≥5.6 mmol/l	6 (24%)	1 (10%)
Triglycerides ≥1.7 mmol/l	5 (20%)	0

Cette analyse a été effectuée en utilisant la méthode des centres les plus proches (Tibshirani, Hastie et al. 2002) en utilisant le package « *pamr* » sous R. Nous obtenons une liste de 42 gènes que nous présentons dans la Table 27.

Table 27 : Sélection des gènes prédicteur Obèses versus Témoins

CloneID	Symbol	GeneID	Chr	Name
2477598	SLPI	6590	20	Secretory leukocyte peptidase inhibitor
2566688	ARL3	403	10	ADP-ribosylation factor-like 3
841070	YARS	8565	1	Tyrosyl-tRNA synthetase
210687	AGTR1	185	3	Angiotensin II receptor, type 1
1841906	TWIST1	7291	7	Twist homolog 1
75254	CSRP2	1466	12	Cysteine and glycine-rich protein 2
1631194	DYNLT1	6993	6	Dynein, light chain, Tctex-type 1
1895613	FN3K	64122	17	Fructosamine 3 kinase
840677	HLA-C	3107	2	Major histocompatibility complex, class I, C
80796	IGL@	3535	22	Immunoglobulin lambda joining 3
235882	MRC2	9902	17	Mannose receptor, C type 2
769948	SERPINB9	5272	6	Serpin peptidase inhibitor, clade B (ovalbumin), member 9
2572139	RPL6	6128	18	Ribosomal protein L6
2546786	RPL41	6171	22	Ribosomal protein L41
2271240	RPS12	6206	6	Ribosomal protein S12

1630974	SULT1A1	6817	16	Sulfotransferase family, cytosolic, 1A, phenol-preferring, member 1
230261	RALA	5898	7	V-ral simian leukemia viral oncogene homolog A (ras related)
183811	TNFSF10	8743	3	Tumor necrosis factor (ligand) superfamily, member 10
782503	FADS1	3992	11	Fatty acid desaturase 1
85394	PPAP2B	8613	1	Phosphatidic acid phosphatase type 2B
81357	KLF9	687	9	Kruppel-like factor 9
461761	ANG	283	14	Angiogenin, ribonuclease, RNase A family, 5
272148	TRPM1	4308	15	Transient receptor potential cation channel, subfamily M, member 1
287327	IGF1	3479	12	Insulin-like growth factor 1 (somatomedin C)
950372	HNRNPH1	3187	5	Heterogeneous nuclear ribonucleoprotein H1 (H)
81601	HLF	3131	17	Hepatic leukemia factor
951008	SC5DL	6309	11	Sterol-C5-desaturase (ERG3 delta-5-desaturase)-like
898076	IRF2BP2	359948	1	Interferon regulatory factor 2 binding protein 2
1590269	SLC2A4	6517	17	Solute carrier family 2 (facilitated glucose transporter), member 4
529307	TOMM7	54543	7	Translocase of outer mitochondrial membrane 7 homolog (yeast)
208993	UXS1	80146	2	UDP-glucuronate decarboxylase 1
162479	ELF3	1999	1	E74-like factor 3 (ets domain transcription factor, epithelial-specific)
68605	IGFBP7	3490	4	Insulin-like growth factor binding protein 7
590150	MT2A	4502	16	Metallothionein 2A
971372	NPTX2	4885	7	Neuronal pentraxin II
884525	FOXJ3	22887	1	Forkhead box J3

510272	CNIH4	29097	1	Cornichon homolog 4 (Drosophila)
1606534	CCDC72	51372	3	Coiled-coil domain containing 72
743860	C17orf79	55352	17	Chromosome 17 open reading frame 79
85194	C5orf26	114915	5	Chromosome 5 open reading frame 26
884514	KIAA1946	165215	2	KIAA1946
489800	SCARA5	286133	8	Scavenger receptor class A, member 5 (putative)

Afin d'explorer les rôles biologiques de cette sélection de gènes discriminants, nous avons réalisé une analyse fonctionnelle à l'aide de l'outil FunNet (Henegar, Tordjman et al. 2008; Prifti, Zucker et al. 2008). Les profils fonctionnels de cette sélection de gènes sont illustrés dans la Fig. 53 pour le système d'annotation KEGG (Kyoto Encyclopedia of Genes and Genomes) et dans la Fig. 54 pour le système Gene Ontology (GO) – Biological Process. Ces profils fonctionnels suggèrent que les gènes les plus informatifs pour distinguer les obèses des témoins normo pondéraux sont liés aux métabolismes lipidiques (« *lipid metabolic proces* ») et protidiques (« *Ribosome* », « *Phenylalanine, tyrosine and tryptophan biosynthesis* », « *translational élongation* »). Des études récentes réalisées dans notre laboratoire (Henegar, Tordjman et al. 2008) ont démontré en effet une altération des fonctions métaboliques du tissu adipeux chez l'obèse caractérisée par une sous expression significative des gènes impliqués dans le métabolisme lipidique, glucidique, protéique et énergétique.

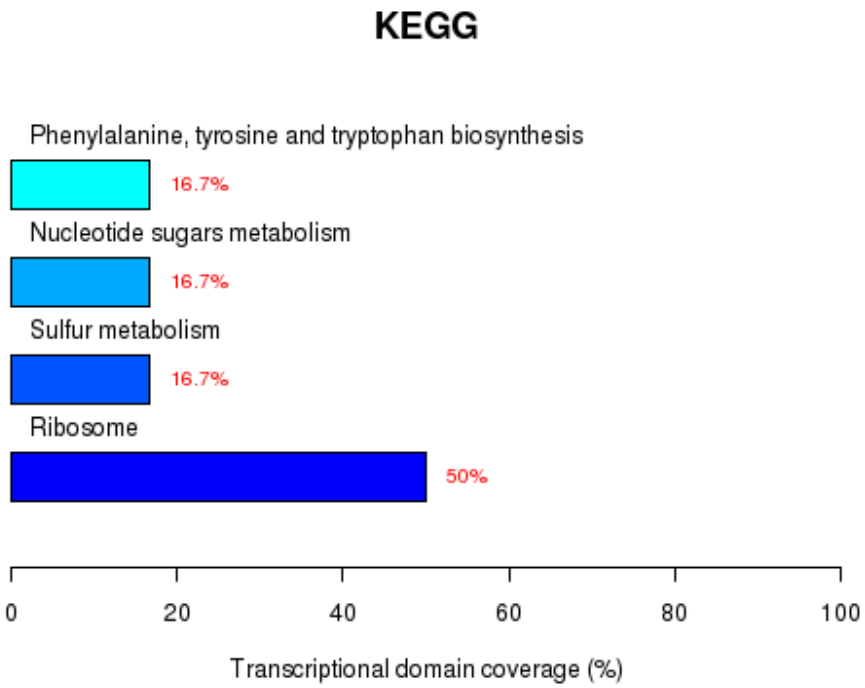


Fig. 53 : Thèmes fonctionnels KEGG significativement surreprésentés parmi ceux annotant la liste des 42 gènes démontrant une expression différentielle significative dans le tissu adipeux des sujets obèses.

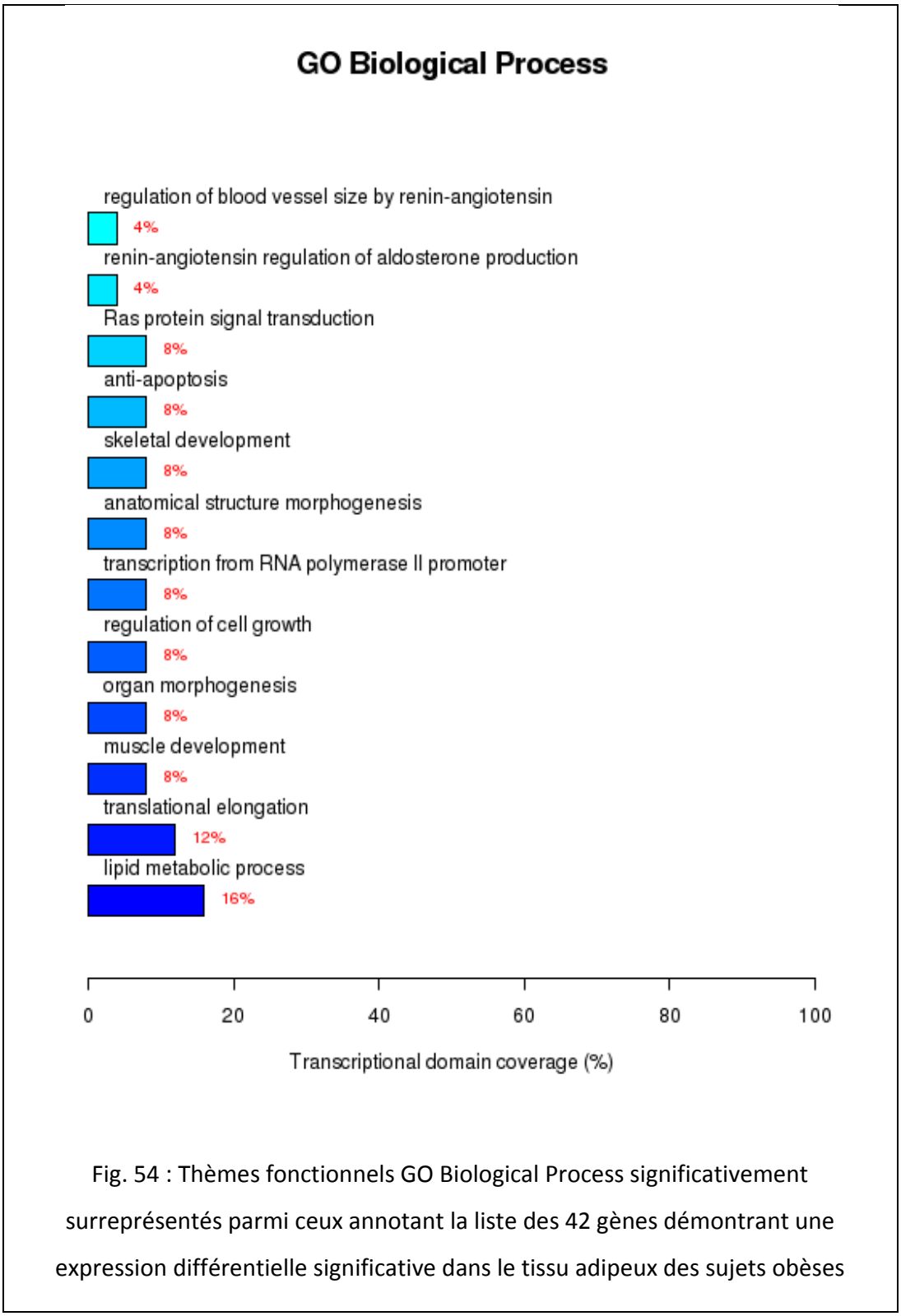


Fig. 54 : Thèmes fonctionnels GO Biological Process significativement surreprésentés parmi ceux annotant la liste des 42 gènes démontrant une expression différentielle significative dans le tissu adipeux des sujets obèses

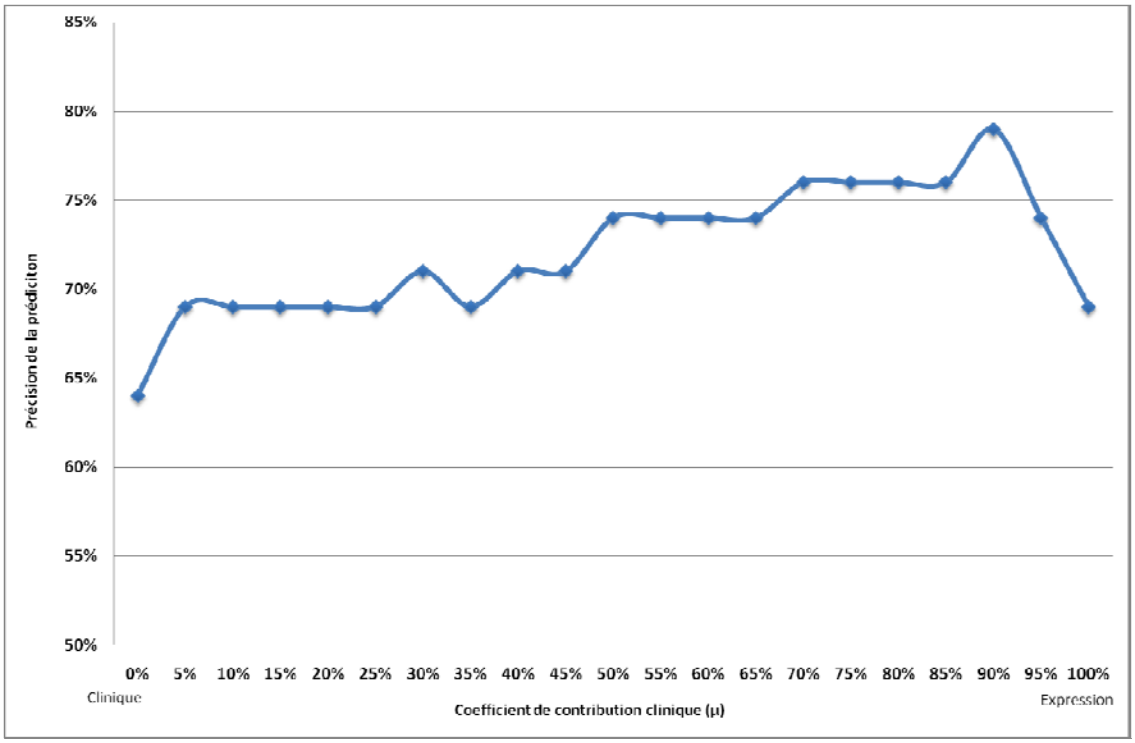


Figure 55 : Variation de la précision de la prédiction (Perte de poids suite à un régime faible en calories) en combinant les données cliniques avec la sélection des gènes obèses versus témoins

La Figure 55, traduit les résultats de la combinaison des données clinique avec la sélection de gènes obèses versus témoins. Nous remarquons une amélioration de la prédiction avec la sélection de gènes. Cette sélection, à elle seule, permet de prédire la perte de poids avec une précision de 69%, soit une amélioration de 16% par rapport à une prédiction à partir de l'ensemble des gènes. Une combinaison de cette sélection avec les données cliniques permet d'obtenir une précision de la prédiction d'environ 80%.

5.1.5.3 **Effet de la randomisation sur la combinaison**

Dans cette partie, nous avons analysé l'effet de la randomisation de chacune des deux sources de données afin d'étudier l'effet qu'aurait la combinaison d'une source randomisée avec une source réelle sur les performances de la prédiction.

Pour le jeu de données du cancer du poumon Harvard, la Figure 56 fournit un comparatif entre la combinaison des données réelles et une combinaison d'une source réelle avec une source permutée. Nous remarquons une altération des performances suite à la combinaison des données d'expression permutées avec les données cliniques réelles. Cependant, la permutation des données clinique induit une augmentation des performances par rapport aux résultats obtenus à partir des données cliniques réelles. Nous observons aussi une augmentation de la précision de la prédiction pour une valeur de μ comprise entre 0.5 et 0.85. Pour une valeur de μ supérieure à 0.85, la précision obtenue à partir des données réelles est supérieure à celle obtenue à partir de la combinaison des données expression avec les données cliniques permutées.

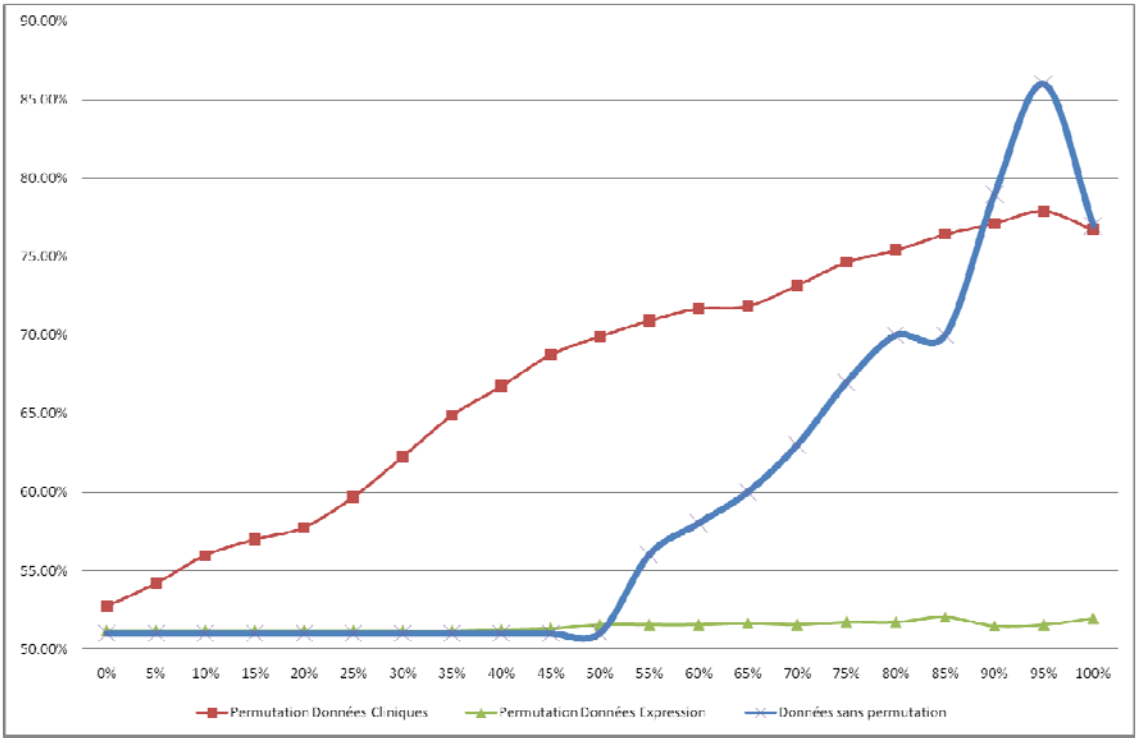


Figure 56 : Effet de la randomisation sur les données du cancer du poumon (survie après 5 ans, données Harvard)

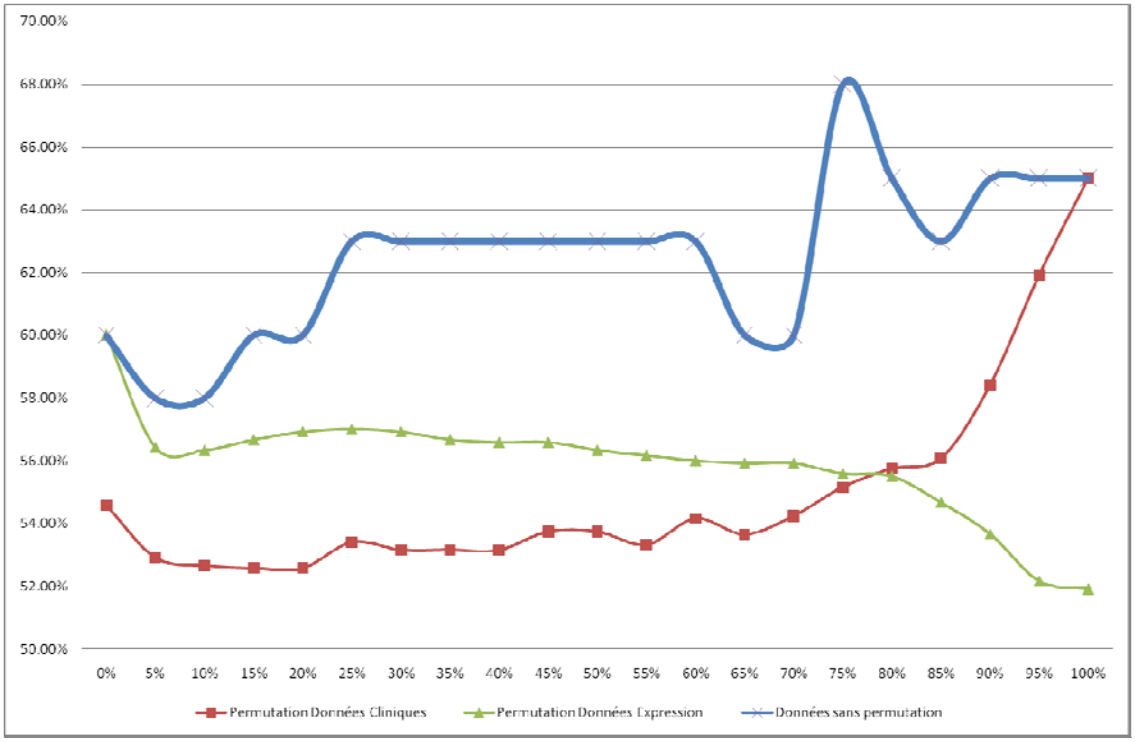


Figure 57 : Effet de la randomisation sur les données du cancer du poumon (survie après 5 ans, données Massachusetts)

Pour le jeu de données du cancer du poumon Massachusetts (Figure 57), nous remarquons une altération des performances suite à la combinaison des données d'expression permutées avec les données cliniques réelles. Ce même phénomène est observé lors de la combinaison des données cliniques permutées avec les données d'expressions.

Pour la base de l'obésité, les résultats de la Figure 58, montrent que pour cette base la permutation des données, que ce soit cliniques ou d'expressions géniques, combinée avec les données réelles dégrade les performances en prédiction du modèle de combinaison.

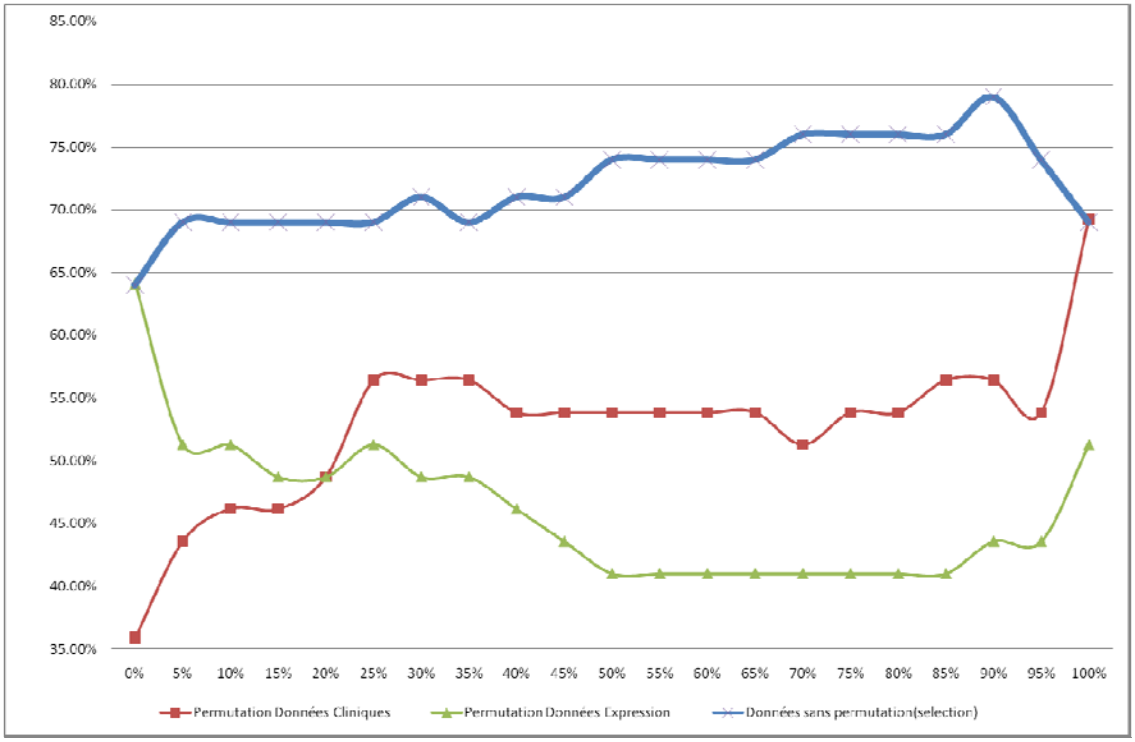


Figure 58 : Effet de la randomisation sur les données de l'obésité VLCD (Perte de poids suite à un régime faible en calories)

5.1.6 **Discussion**

Nous avons présenté ici une méthode de combinaison qui agrège les similarités calculées à partir des données cliniques et transcriptomiques de sorte à obtenir une mesure globale capable de donner un indicateur plus performant de la prédiction. L'analyse des bases utilisées montre une supériorité informative des données de puce à ADN par rapport aux données cliniques, ceci rejoint les résultats retrouvés dans plusieurs analyses prédictives dans le domaine de l'oncologie (Cardoso 2003; Wong, Selvanayagam et al. 2003; Harima, Togashi et al. 2004; Selvanayagam, Cheung et al. 2004; Bidus, Risinger et al. 2006; Watanabe, Komuro et al. 2006; Morrow 2007; Watanabe, Kobunai et al. 2007; Gevaert, Van Vooren et al. 2008; Rimkus, Friederichs et al. 2008). Cependant, comme nous avons pu le montrer expérimentalement, les données cliniques combinées avec les données transcriptomiques améliorent les performances des modèles prédictifs. La visualisation de la variation de la précision de la prédiction en

fonction de la pondération entre les données cliniques et les données transcriptomiques donne une idée de l'importance des sources. Cependant, avant de tirer des conclusions et afin de s'assurer de la fiabilité des résultats, il est important de lancer des tests de permutation pour vérifier que l'amélioration obtenue n'est pas la conséquence d'un artefact. L'exploitation d'autre expérimentation s'avère intéressante et peut amener à trouver des gènes cibles qui pourraient faire partie des marqueurs potentiels de la signature de l'obésité. Cette piste est à creuser davantage, en effet l'intégration des connaissances du domaine et l'exploitation d'autres expérimentations pourraient favoriser la compréhension des mécanismes génétiques découverte de relation et rendre plus transparente les relations qui existent entre les différentes listes de gènes cibles. Ces informations apportent des connaissances pertinentes pour les biologistes et les cliniciens pour mieux répondre à certaines questions autour de l'obésité.

Nous avons constaté à partir de nos analyses et de la littérature qu'il existe un nombre non négligeable de bases pour lesquelles les résultats en prédiction sont faibles. Cela peut être dû à la difficulté à prédire certaines instances présentes dans la base. Pour ces cas-là, nous pouvons considérer des modèles qui se servent d'un indice de fiabilité pour décider de la réponse à donner à ces instances avec la possibilité de s'abstenir sur les cas avec un faible indice de fiabilité. C'est ce que nous présentons dans la section suivante.

5.2 Combinaison de modèle d'apprentissage à partir des données biopuces

Dans la vie courante, il existe des situations où la prise de décisions est un peu délicate. Par exemple, un médecin généraliste confronté à un cas clinique inhabituel et ambigu peut ne pas se prononcer sur la maladie de ce patient et l'envoyer consulter un spécialiste. De même, dans les élections, certains candidats préfèrent s'abstenir que de voter un candidat non apprécié. Malgré un usage courant de la pratique de l'abstention dans la vie courante, ce principe est rarement utilisé en apprentissage automatique car le choix de s'abstenir se base sur

une multitude de facteurs qui sont difficile à quantifier. Pour comprendre le problème, considérons l'exemple suivant.

On suppose qu'on dispose de deux oracles qu'on peut consulter pour la prise de décision. Le premier donne une réponse à n'importe quelle question, peu importe le niveau de confiance qu'il a à propos de la réponse. Le second à la possibilité de ne pas se prononcer sur une question quand il n'est pas très confiant. La question qui se pose : À quel oracle doit-on faire confiance ?

En fait, tout dépend du nombre de bonnes réponses données par l'oracle. Prenons le cas du premier oracle, soit il est omniscient et sait réellement répondre à toutes les questions ou bien il est incapable d'admettre qu'il existe des réponses qu'il ignore, ce qui est plus probable. Dans la seconde hypothèse, son avis échoue dans de nombreux cas et les gens vont se tourner vers le deuxième oracle qui ne répond que lorsqu'il a un degré de confiance « satisfaisant » en sa réponse. Mais là encore, la manière utilisée par l'oracle pour fixer le seuil de confiance qui lui dicte s'il faut répondre ou non est importante. Si le seuil est très élevé, l'oracle aura tendance à s'abstenir dans la plupart des cas et dans ce cas les gens vont se retourner de nouveau vers le premier oracle, car ils ont rarement une réponse du second oracle. Par conséquent, un bon oracle est celui qui arrive à trouver un bon compromis entre ces deux situations extrêmes.

Cet exemple illustre les différents problèmes rencontrés lors de l'extension de la classification standard à une classification avec abstention. De toute évidence, un classeur avec abstention semble plus intéressant qu'un classeur standard. Cependant, il faut trouver un bon compromis entre le taux d'abstention et la précision de la prédiction.

5.2.1 **Classeur avec abstention**

Les modèles d'apprentissage avec abstention, appelé aussi **« classeurs prudents »** (Ferri and Hernandez-Orallo 2004), sont des classeurs capables de s'abstenir sur les réponses avec une faible confiance.

Table 28 : Matrice de Confusion d'un classeur

	Classe prédite	
Vrai Classe	**P**	**N**
P	C(P, p)	C(P, n)
N	C(N, p)	C(N, n)

Le classeur avec abstention étend l'ensemble de classe originale C en ajoutant une classe supplémentaire $\perp$, celle des éléments de faible confiance.

Table 29: matrice de Confusion d'un classeur avec abstention

	Classe Prédite		
Vrai Classe	**P**	**n**	$\perp$
P	C(P, p)	C(P, n)	C(P, $\perp$)
N	C(N, p)	C(N, n)	C(N, $\perp$)

Il existe différentes mesures de performances basées sur les matrices de confusion qui permettent de calculer des fonctions de coût. L'efficacité et la capacité sont deux mesures de performances des modèles d'apprentissage avec abstention qui correspondent à des aires de surface dans un espace bidimensionnel. Soit l'espace (précision, α) pour l'efficacité et (abstention, α) pour la capacité(Ferri and Hernandez-Orallo 2004).Ainsi, les performances d'un modèle abstinent peuvent être obtenues en se basant sur la matrice de confusion et la matrice de coût.

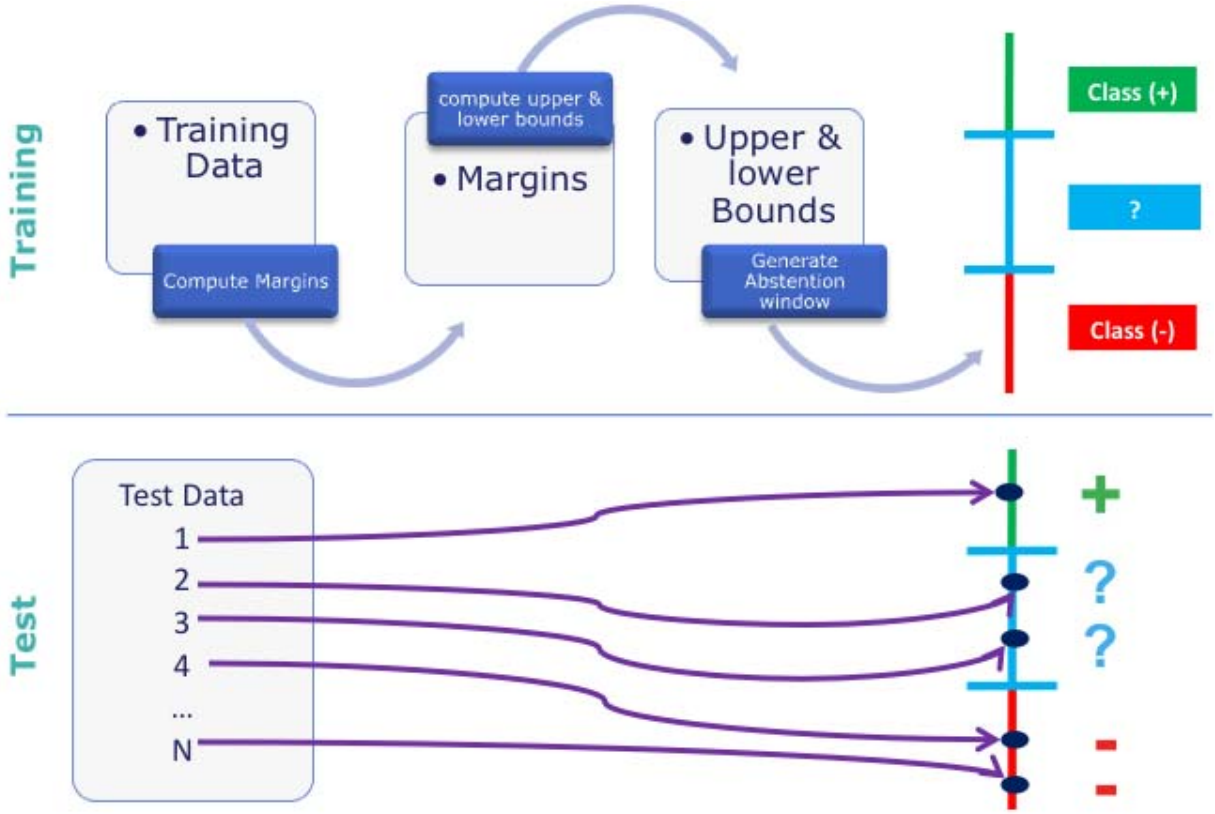


Figure 59 : principe de la classification avec abstention

5.2.1.1 **Principe de l'approche de classification avec abstention**

L'approche que nous avons adoptée est inspirée des travaux de Friedel et al. (Friedel 2005), nous repérons dans ce qui suit l'idée générale et le principe de fonctionnement, le lecteur est amené à consulter (Friedel 2005; Friedel, Ulrich et al. 2006) pour plus d'approfondissement. Un méta-classeur s'abstient sur les instances de faible confiance. La confiance est basée sur un score fourni par le classeur de base. En effet, ce score peut être une probabilité de classe pour l'algorithme de bayes par exemple ou bien une distance de l'hyperplan de séparation pour les SVMs. Le degré de confiance est basé sur la différence de score entre les deux classes, la marge.

5.2.1.2 **Calcul de la marge**

Soit $s_p(x)$ le score pour la prédiction positive d'une instance $x \in \chi$ et $s_n(x)$ le score pour la prédiction négative. La *marge* d'une instance est définie comme étant la différence entre ces deux scores $m(x) = s_p(x) - s_n(x)$. Une instance est étiquetée positive si $m(x)$ est positive et celle-ci est négative dans le cas contraire. En plus d'être utilisée pour prédire la classe d'une instance, la marge donne un indicateur sur la fiabilité de cette prédiction. Si la valeur absolue de la marge est grande, nous pouvons prédire avec une confiance plus élevée la classe prédite qu'avec une marge de petite valeur. Voici un exemple où le score reflète une probabilité de classe

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
$S_p(X_i)$	0.05	0.3	0.55	0.2	0.65	0.9	0.35	0.4	0.7	0.6
$S_n(X_i)$	0.95	0.7	0.45	0.8	0.35	0.1	0.65	0.6	0.3	0.4
$m(X_i)$	-0.9	-0.4	0.1	-0.6	0.3	0.8	-0.3	-0.2	0.4	0.2
Y_i	N	N	N	N	N	P	P	P	P	P

Table 30: Exemple de classe de probabilité avec 5 instances négatives et 5 instances positives

Dans cet exemple, les instances x1 et X6 sont correctement classées avec une confidence élevée, alors que les instances X3 et X10 ont la plus petite marge (en valeur absolue). Pour cet exemple le taux de précision est de 60%. Pour ce cas le critère de décision pour l'attribution de la classe étant le signe de la marge $m(x)$ c.à.d. si $m(x)$ est positive alors l'algorithme attribut la classe « P » et dans le cas contraire, la classe « N ». Maintenant, si on applique une restriction sur la valeur de la marge de telle manière à ne garder que les instances X ayant une marge $m(x) < -0.25$ ou $m(x) > 0.15$ par exemple, nous observons que cette restriction sur les valeurs a faible marge nous permet d'avoir un gain de performances et la précision passe à 75%. Cet exemple introduit l'idée intuitive d'une fenêtre d'abstention afin d'améliorer la fiabilité de la prédiction.

5.2.1.3 **Fenêtre d'abstention**

La fenêtre d'abstention est définie par un couple (l,u) tel que la prédiction d'une instance $x \in \chi$ est donnée par $\pi(a,x) = \begin{cases} p & si \quad m(x) \geq u \\ \perp & si \quad l < m(x) < u \\ n & si \quad m(x) \leq l \end{cases}$

la construction d'un classeur abstinent requiert deux phases d'apprentissage séparées. Dans la première phase, un classeur standard est construit à partir d'un ensemble d'apprentissages et ensuite applique à un ensemble de validation V_n pour récupérer les marges $M = (m(x_1),(x_2),(x_3),.....,m(x_n))$. Ces marges vont être utilisées par la suite pour calculer la valeur optimale du couple (l,u) qui désigne les frontières de la fenêtre d'abstention.

5.2.2 **Modèles d'apprentissage abstinent avec délégation**

Les modèles d'apprentissage avec abstention ne traitent pas les instances sur lesquelles on s'abstient. Une approche intéressante qui pourrait réduire le taux d'abstention tout en essayant d'améliorer, ou du moins maintenir, de bonnes performances de ces modèles. Ferri et al. proposent une approche de délégation ou un premier classeur construit à partir de l'intégralité du jeu de données renvoie les instances à faible confiance à un deuxième classeur. Le principe de la délégation s'inspire en quelques sortes de l'approche « diviser pour régner » et plus particulièrement de l'algorithme PART (Frank and Witten 1998) dont le principe consiste à apprendre un arbre de décision, sélectionne la branche qui a la couverture maximale, retire le reste de l'arbre et sélectionne un second arbre avec le reste des exemples. Le processus est itéré jusqu'à ce que tous les exemples soient couverts. La délégation est indépendante du classeur utilisé, elle se base sur un seuil de confiance qui dans le cas des arbres de décision par exemple pourrait sélectionner plusieurs branches avec une grande couverture.

La délégation est une méthode multi-classeur qui opère en série (et non pas en parallèle ou d'une manière hiérarchique), elle se base sur le transfert (pas de combinaison) et préserve les attributs. Les avantages de la délégation sont multiples. Tout d'abord, le résultat global de classification n'est pas une combinaison des classeurs, mais une liste de décision, si on utilise les

arbres de décisions comme étant la base du classement, la classification est un arbre de décision, et ses décisions peuvent être retrouvées et comprises assez facilement. De plus, étant donné que certaines parties des modèles ne seront pas utilisées, il est possible, dans certains cas, de simplifier les modèles, par exemple, dans le cas des arbres de décision, un sous arbre peut être taillé si toutes ses feuilles mènent à la délégation.

Voici une brève comparaison du classement avec délégation avec d'autres techniques utilisées pour améliorer la précision de prédiction. Les méthodes de délégation qui sont voisines des approches ensemblistes telles que le boosting (Freund and Schapire 1999) et le stacking (Wolpert 1992) sont plus performantes que les approches de combinaison, en général (Ferri, Flach et al. 2004). Le boosting attribue un poids (coût) élevé pour les exemples incorrectement classés et un poids faible pour les exemples qui sont classés correctement à chaque itération. La délégation, d'autre part, supprime les exemples qui sont classés avec une grande confiance et laisse les exemples qui sont classés avec faible confiance pour les itérations qui suivent. Le stacking construit dans une deuxième étape un méta-classeur qui décide quel classeur de base utiliser (à partir d'un ensemble indépendant de classeurs). D'autres méthodes de classification multiple, comme la *cascade generalization* (Gama and Brazdil 2000) génère de nouveaux attributs à partir de l'estimation de probabilité des classes données par les classeurs de base ou par la subdivision l'arbre de décision. En revanche, la délégation a produit des modèles qui sont entièrement et exclusivement définis par les attributs d'origine et la classe. L'arbitrage (Ortega, Koppel et al. 2001) (Ortega et al., 2001) et le *grading* (Seewald, F\ et al. 2001) sont aussi liés à la délégation, mais les deux arbitres extérieurs apprennent à évaluer la probabilité d'erreur de chaque classeur à partir d'un ensemble de classeurs de base et leurs domaines d'expertise .

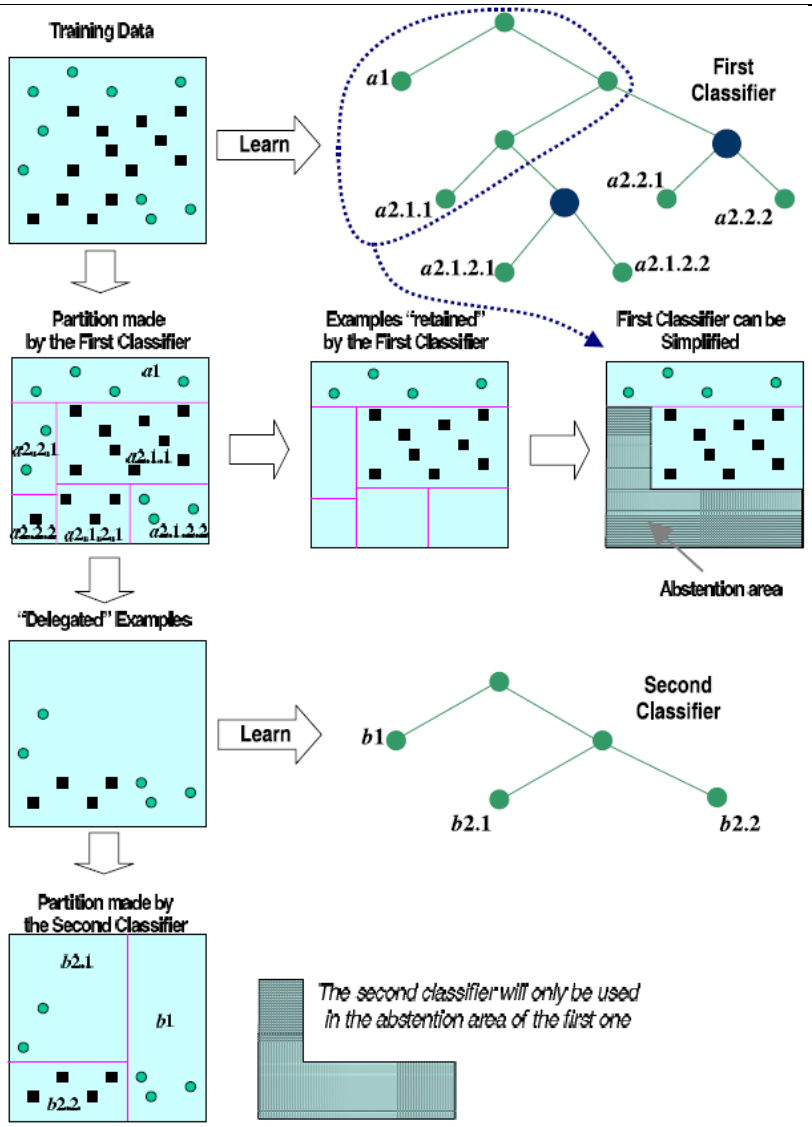


Figure 60 : modèle de délégation avec les arbres de décision (Ferri, Flach et al. 2004)

5.2.3 Modèle d'apprentissage avec Abstention/délégation pour l'apprentissage à partir de sources multiples

L'approche d'abstention/délégation implique la construction de deux modèles : un premier modèle $\psi 1$ sera construit à partir des données biopuce et un sous-modèle $\psi 2$ à partir des données cliniques. Si le modèle $\psi 1$ prédit le résultat d'un patient avec une confiance élevée, dans ce cas-là, la classe prédite par le modèle est attribuée à cette instance. Dans le cas

contraire où la confiance est faible, (1 délègue l'instance à ψ2 qui a son tour aura la possibilité d'attribuer une classe dans le cas d'une confiance élevée ou bien s'abstenir dans le cas contraire. Une fois que toutes les instances sont testées, le modèle est évalué sur la base de la précision de la prédiction, mais aussi le taux d'abstention.

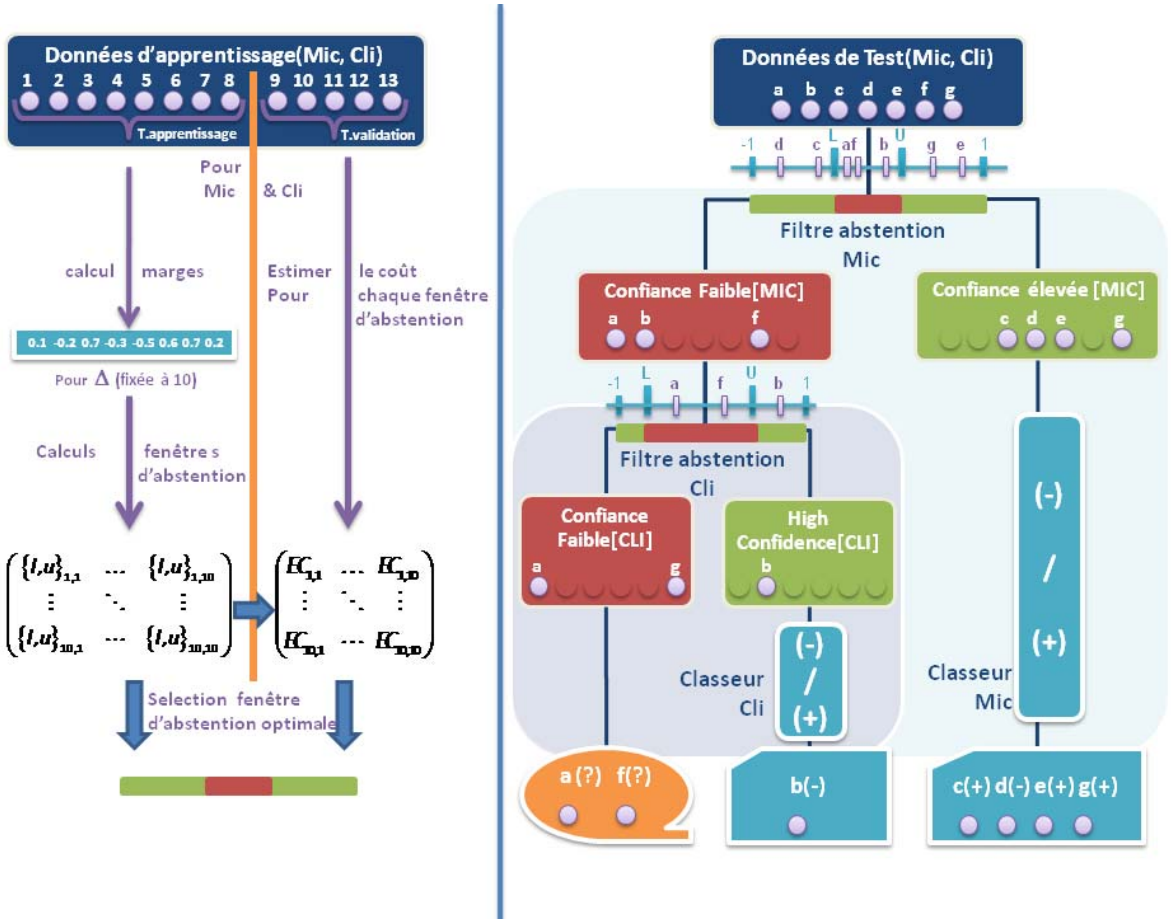


Figure 61: modèle d'abstention/délégation à partir des données cliniques et biopuces.

Comme dans le cadre de l'apprentissage classique, l'apprentissage avec délégation/abstention est une construction en deux phases. Dans la phase d'apprentissage, les instances sont divisées aléatoirement en un groupe d'apprentissage et un groupe de test selon le procédé de validation croisée, avec un ratio de 90% pour l'apprentissage et 10% pour le test. Pour chaque instance deux sources de données (cliniques et microarray) sont disponibles. En un premier temps, les données d'expression sont utilisées pour générer le modèle ψ1 à partir des

instances d'apprentissage, ce qui permet d'obtenir la fenêtre d'abstention optimale $\varphi 1$ calculé à partir des probabilités à priori générées par les classeurs classiques pour chaque instance d'apprentissage. Ensuite, les probabilités des instances de test de ce modèle sont calculées et leurs valeurs détermineront le degré de confiance par rapport à leur appartenance ou pas à la fenêtre d'abstention $\varphi 1$. Les instances se situant en dehors de la fenêtre d'abstention $\varphi 1$ sont les instances retenues par le classeur et par conséquent il leur est attribué une classe par le classeur. Les instances se situant dans la fenêtre d'abstention ne reçoivent pas de classe et sont passées au deuxième modèle $\psi 2$ qui, à son tour, va construire une deuxième fenêtre d'abstention $\varphi 2$ et décidera de la même manière entre s'abstenir sur les instances restantes ou bien leurs attribuer une classe. Idéalement, le modèle $\psi 2$ a pour but de réduire le taux d'abstention de $\psi 1$.

5.2.4 **Résultats**

Dans cette analyse, nous avons évalué les performances de cinq algorithmes: SVM, Random Forest, naïf de Bayes, et l'arbre de décision(J4.8). Nous avons comparé la précision de prédiction en classification des modèles standard à notre proposition d'abstention / délégation. Nous rapportons à chaque fois la précision et le taux d'abstention, qui définissent les performances du modèle.

5.2.4.1 **Données cancer**

Nous présentons ci-après, les résultats obtenus en utilisant notre méthode d'abstention délégation sur les données du cancer. La *Table 31* résume les résultats de la prédiction de la survie après 5 ans à partir de la base Harvard dans les différents scénarios : sans abstention, avec abstention, avec délégation. Pour chaque scénario, nous donnons la précision de la prédiction ainsi que le taux d'abstention.

Données Harvard		Précision (%)				
		SVM	Random Forests	Naive Bayes	PART	J4.8
sans Abstention	Précision (Clinical)	75.00%	77.50%	72.50%	70.00%	67.50%
	Précision (Microarray)	72.50%	55.00%	67.50%	52.50%	50.00%
Abstention Clinical	Précision	78.57%	79.49%	93.10%	70.00%	64.86%
	Taux D'abstention	30.00%	2.50%	27.50%	0.00%	7.50%
Abstention Microarray	Précision	72.50%	55.00%	67.50%	51.28%	50.00%
	Taux D'abstention	0.00%	0.00%	0.00%	2.50%	0.00%
délégation (Cli -> Micro)	Précision	77.50%	81.63%	90.00%	75.51%	67.35%
	Taux D'abstention	0.00%	0.00%	0.00%	0.00%	0.00%
délégation (Micro -> Cli)	Précision	72.50%	55.00%	67.50%	52.50%	50.00%
	Taux D'abstention	0.00%	0.00%	0.00%	0.00%	0.00%

Table 31 : Précision de la prédiction (survie après 5 ans, données Harvard)

Pour la base Harvard, les précisions des algorithmes standards varient entre 50% et 77.5%.Le meilleur résultat est obtenu par les random forest avec les données cliniques. Avec l'abstention, nous obtenons des précisions variant entre 50% et 93.1%. Le meilleur résultat est obtenu avec l'algorithme de bayes appliqué aux données cliniques et pour cette précision nous avons un taux d'abstention de 27.50%. Avec la délégation la précision de la prédiction varie entre 50% et 90%. Le meilleur résultat est là aussi obtenu par l'algorithme de bayes et toutes les instances déléguées du modèle clinique au modèle microarray ont été prédites et donc le taux d'abstention est nul.

Donnée Michigan		Précision (%)				
		SVM	Random Forests	Naive Bayes	PART	J4.8
sans Abstention	Précision (Clinical)	62.22%	64.44%	58.89%	64.44%	62.22%
	Précision (Microarray)	42.22%	63.33%	67.78%	55.56%	60.00%
Abstention Clinical	Précision	86.36%	62.20%	70.15%	64.47%	63.16%
	Taux D'abstention	51.11%	8.89%	25.56%	15.56%	15.56%
Abstention Microarray	Précision	41.57%	62.65%	69.41%	55.56%	60.23%
	Taux D'abstention	1.11%	7.78%	5.56%	0.00%	2.22%

délégation (Cli -> Micro)	Précision	63.53%	61.11%	67.44%	63.75%	61.45%
	Taux D'abstention	1.16%	1.10%	1.15%	1.23%	2.35%
délégation (Micro -> Cli)	Précision	42.22%	63.53%	69.77%	56.04%	59.34%
	Taux D'abstention	0.00%	5.56%	1.11%	0.00%	0.00%

Table 32 : Précision de la prédiction (survie après 5 ans, données Michigan)

Avec les données Michigan (*Table 32*), les précisions des algorithmes standards varient entre 42.2% et 67.8%. Le meilleur résultat est obtenu par l'algorithme de bayes avec les données biopuces. Avec l'abstention, nous obtenons des précisions variant entre 41.6% et 86.4%. Le meilleur résultat est obtenu avec les SVM appliqués aux données cliniques et pour cette précision nous avons un taux d'abstention de 51.1%. La délégation n'apporte pas d'amélioration aux résultats sur cette base.

Avec les données Massachusetts (Table 33), les précisions des algorithmes standards varient entre 37.5% et 75%. Le meilleur résultat est obtenu par l'algorithme Part et J4.8 avec les données biopuces. Pour cette base, l'abstention ainsi que la délégation n'apportent pas d'amélioration aux résultats sur cette base du fait que les instances se trouvent en dehors de la fenêtre d'abstention dans la plupart des cas.

Données MIT		Précision (%)				
		SVM	Random Forests	Naive Bayes	PART	J4.8
sans Abstention	Précision(Clinical)	37.50%	42.50%	45.00%	37.50%	37.50%
	Précision(Microarray)	67.50%	47.50%	50.00%	75.00%	75.00%
Abstention Clinical	Précision	35.29%	42.11%	46.15%	39.47%	36.11%
	Taux D'abstention	15.00%	5.00%	2.50%	5.00%	10.00%
Abstention Microarray	Précision	69.23%	47.50%	50.00%	75.00%	75.00%
	Taux D'abstention	2.50%	0.00%	2.44%	2.44%	2.44%
délégation (Cli -> Micro)	Précision	42.50%	42.50%	45.00%	42.50%	40.00%
	Taux D'abstention	0.00%	0.00%	0.00%	0.00%	0.00%
délégation (Micro -> Cli)	Précision	67.50%	47.50%	50.00%	75.00%	75.00%
	Taux D'abstention	0.00%	0.00%	0.00%	0.00%	0.00%

Table 33 : Précision de la prédiction (survie après 5 ans, données Massachusetts)

5.2.4.2 **Données obésité**

Nous présentons ci-après, les résultats obtenus en utilisant notre méthode d'abstention délégation sur les données de l'obésité.

La Table 34 résume les résultats de la prédiction de la Perte de poids suite à un régime faible en calories à partir des données Nugenob dans les différents scénarios : sans abstention, avec abstention, avec délégation. Pour chaque scénario, nous donnons la précision de la prédiction ainsi que le taux d'abstention. Pour cette base, les précisions des algorithmes standards varient entre 36% et 68%. Le meilleur résultat est obtenu par les SVM avec les données cliniques. Pour cette base, l'abstention ainsi que la délégation n'apportent pas d'amélioration aux résultats sur cette base du fait que les instances se trouvent en dehors de la fenêtre d'abstention dans la plupart des cas.

Données Nugenob		Précision (%)				
		SVM	Random Forests	Naive Bayes	PART	J4.8
sans Abstention	Précision (Clinical)	54.00%	36.00%	40.00%	40.00%	38.00%
	Précision (Microarray)	68.00%	56.00%	50.00%	50.00%	50.00%
Abstention Clinical	Précision	52.17%	36.17%	51.43%	40.00%	36.17%
	Taux D'abstention	8.00%	6.00%	30.00%	0.00%	6.00%
Abstention Microarray	Précision	65.22%	55.32%	51.61%	47.83%	47.83%
	Taux D'abstention	8.00%	6.00%	38.00%	8.00%	8.00%
délégation (Cli -> Micro)	Précision	56.00%	41.18%	41.67%	42.31%	38.00%
	Taux D'abstention	0.00%	4.00%	0.00%	0.00%	0.00%
délégation (Micro -> Cli)	Précision	65.22%	56.25%	47.50%	48.94%	48.94%
	Taux D'abstention	0.00%	4.00%	9.09%	0.00%	0.00%

Table 34 : Précision de la prédiction (Perte de poids suite à un régime faible en calories, données Nugenob)

Avec les données VLCD, la précision de la prédiction varie entre 36% et 70%. Les précisions des algorithmes standards varient entre 37.5% et 75%. Le meilleur résultat est obtenu par l'algorithme Part et J4.8 avec les données cliniques. Pour cette base, l'abstention ainsi que la délégation n'apportent pas d'amélioration aux résultats sur cette base du fait que les instances se trouvent en dehors de la fenêtre d'abstention dans la plupart des cas.

Données VLCD		Précision (in%)				
		SMO	Random Forests	Naive Bayes	PART	J4.8
sans Abstention	Précision (Clinical)	63.33%	60.00%	43.33%	70.00%	70.00%
	Précision (Microarray)	60.00%	63.33%	53.33%	43.33%	43.33%
Abstention Clinical	Précision	70.00%	58.62%	36.00%	70.00%	70.00%
	Taux D'abstention	33.33%	3.33%	16.67%	0.00%	0.00%
Abstention Microarray	Précision	60.71%	62.96%	64.00%	43.33%	48.15%
	Taux D'abstention	6.67%	10.00%	16.67%	0.00%	10.00%
délégation (Cli -> Micro)	Précision	66.67%	60.00%	42.86%	70.00%	70.00%
	Taux D'abstention	0.00%	0.00%	3.45%	0.00%	0.00%
délégation (Micro -> Cli)	Précision	62.07%	62.07%	63.33%	43.33%	46.67%
	Taux D'abstention	0.00%	3.33%	0.00%	0.00%	0.00%

Table 35 : Précision de la prédiction (Perte de poids suite à un régime faible en calories, données internes)

5.2.5 **Discussion**

Dans cette partie, nous avons proposé une nouvelle approche d'abstention/délégation qui vise à exploiter plusieurs sources afin de fournir un modèle précis et performant. Pour les bases de l'obésité, cette approche se montre peu efficace, mais cela rejoint les précédents résultats pour confirmer la complexité de la tâche d'apprentissage à partir de ces bases.

Les résultats expérimentaux montrent que par rapport aux algorithmes d'apprentissage standard, l'utilisation de l'abstention peut s'avérer d'intérêt pour améliorer la précision de la

prédiction dans certains cas. Mais l'introduction de cette nouvelle approche donne une nouvelle dimension de la prédiction qui tient en compte non seulement la précision, mais aussi le taux d'abstention. Le choix d'un bon compromis entre ces deux paramètres est subjectif, il dépend de la spécificité du domaine et du cout éventuel d'un mauvais classement.

Conclusion

Au cours de ces travaux de recherche, nous avons analysé l'apport de la bioinformatique à la prédiction de la perte de poids suite à un régime ou une intervention chirurgicale.

Nous avons présenté deux études de la prédiction de la perte de poids suite à une restriction calorique. Nous avons pu constater à travers ces deux analyses une différence significative entre les résultats obtenus. Dans la première analyse, les prédicteurs ont des performances limitées alors que dans la deuxième les modèles de prédiction sont assez performants comparés aux premiers, ce qui est tout de même surprenant. En collaboration avec les biologistes de notre équipe, nous avons essayé de comprendre les raisons qui pourraient être à l'origine d'une telle différence dans les résultats. Nous avons pu écarter certaines hypothèses susceptibles d'être à l'origine de cette variation dans les résultats, par contre le problème reste ouvert par manque d'éléments capables de nous en dire plus sur cette différence. Les bons résultats obtenus dans le cadre du projet Diogenes nous ont encouragés à pousser l'analyse exploratoire et étudier une liste de gène prédicteurs qui apporte des connaissances nouvelles autour du problème de l'obésité.

Dans le cadre de la prédiction de la perte de poids suite à la chirurgie gastrique, nous avons vu que la précision de la prédiction de la perte de poids est très limitée avec les approches classiques. Nous avons alors introduit un modèle de classement hybride, associant une méthode de partitionnement et une méthode de classement qui sont respectivement, les cartes topologiques et les SVMs. Ce modèle utilise l'organisation des données fournie par les cartes topologiques mixtes pour subdiviser l'espace des données afin d'apprendre un SVM spécifique pour chaque sous-espace des données. Cette approche a permis d'améliorer légèrement la précision et réduire la variance des résultats. À l'aide des cartes topologiques, les médecins ont identifié des profils de patients intéressants qui leurs ont permis de mieux comprendre les comportements métaboliques chez certains groupes de patients.

Nous avons élargi notre étude pour s'intéresser à l'étude de l'amélioration de l'état de santé des patients suite à un régime et ceci en regardant le profil d'évolution des paramètres biocliniques. Nous avons constaté que l'évolution de certaines variables comme la glycémie et le triglycéride suite à une intervention de type bypass est plus prédictible que l'évolution de variables comme le poids ou l'IMC. Malgré une précision de la prédiction dépassant les 90%, ses modèles de prédiction sont peu informatifs d'un point de vu biologique. En effet, la prédiction du la variation de la glycémie se fait à partir de glycémie préopératoire uniquement et il en est de même pour la variation du triglycéride. Pour cette intervention chirurgicale, le suivie de l'évolution du profil de santé des patients à long terme est très important. De ce fait, la prédiction du profil d'évolution temporelle est plus importante qu'une prédiction ponctuelle. L'évolution de la base de données de la chirurgie nous permettra de faire une analyse plus robuste des profils d'évolution à long terme, puisque c'est l'amélioration des paramètres biocliniques qui intéresse les médecins, et la connaissance de l'influence d'une intervention chirurgicale sur l'amélioration de la santé des patients obèses, serait d'une grande utilité en clinique.

Les études ont montré une différence quant à la contribution et l'apport des données biocliniques et les données transcriptomiques. Cette hétérogénéité des données peut être exploitée pour l'amélioration de la précision des modèles et aussi pour une meilleure compréhension de la contribution de chaque type de données. Nous avons présenté deux approches de combinaison. La première agrège les similarités calculées à partir des données cliniques et transcriptomique de sorte à obtenir une mesure globale capable de donner un indicateur plus performant de la prédiction. Par cette approche, la combinaison des données cliniques avec les données transcriptomiques a amélioré les performances des modèles prédictifs. La visualisation de la variation de la précision de la prédiction en fonction de la pondération entre les données cliniques et les données transcriptomiques nous a permis d'avoir une idée sur l'importance des sources. Cependant, avant de tirer des conclusions et afin de s'assurer de la fiabilité des résultats, il est important de lancer des tests de permutation pour vérifier que l'amélioration obtenue n'est pas la conséquence d'un artefact. La deuxième approche d'abstention/délégation améliore les résultats pour les bases du cancer mais se

montre peu efficace sur les bases de l'obésité. Les résultats expérimentaux montrent que par rapport aux algorithmes d'apprentissage standard, l'utilisation de l'abstention peut s'avérer d'intérêt pour améliorer la précision de la prédiction. Cette nouvelle approche donne une nouvelle dimension de la prédiction qui tient en compte non seulement la précision, mais aussi le taux d'abstention et le choix d'un bon compromis entre ces deux paramètres dépendant de la spécificité du domaine.

Pour cette deuxième partie relative à la combinaison, le nombre limité de bases pour lesquelles nous disposons à la fois des données transcriptomiques et des données cliniques ne nous a pas permis de faire des analyses avancées. Nous allons nous intéresser dans le futur à la combinaison de groupe de gènes en regardant leurs interactions ainsi que leurs fonctions. Pour construire ces groupes de gènes, nous allons utiliser l'outil FunNet (http://www.funnet.info/) développé dans notre équipe. Cette combinaison nous permettra d'exploiter les informations relatives à l'importance biologique et fonctionnelle des gènes pour améliorer la prédiction.

Un autre aspect limitant est le nombre de sujets inclus dans nos études transcriptomiques et qui ne dépasse pas les 50 sujets. Cette limitation sera dépassée dans le futur, car les analyses à venir du projet Diogènes comporteront des études avec au moins 100 sujets par étude.

Nous allons aussi étendre les méthodes pour qu'elles s'adaptent à un apprentissage à partir de sources multiples puisque dans la suite du projet, en plus des données bioclinqiues et transcriptomiques, des données peptidomiques et métabolomiques seront disponibles.

Nous envisageons aussi de préparer un package R à intégrer à la suite bioconductor et qui sera dédié aux outils bioinformatiques de prédiction à partir de sources multiples.

Bibliographie

Alberg, A. J. and J. M. Samet (2003). "Epidemiology of lung cancer." Chest **123**(1 Suppl): 21S-49S.

Alizadeh, A. A., M. B. Eisen, et al. (2000). "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling." Nature **403**(6769): 503-511.

Ashburner, M., C. A. Ball, et al. (2000). "Gene Ontology: tool for the unification of biology." Nat Genet **25**(1): 25-29.

Bader, G. D., D. Betel, et al. (2003). "BIND: the Biomolecular Interaction Network Database." Nucl. Acids Res. **31**(1): 248-250.

Ball, C. A., I. A. B. Awad, et al. (2005). "The Stanford Microarray Database accommodates additional microarray platforms and data formats." Nucl. Acids Res. **33**(suppl_1): D580-582.

Barrett, T., D. B. Troup, et al. (2007). "NCBI GEO: mining tens of millions of expression profiles--database and tools update." Nucl. Acids Res. **35**(suppl_1): D760-765.

Basdevant, A. (2000). "[Obesity: epidemiology and public health]." Ann Endocrinol (Paris) **61 Suppl 6**: 6-11.

Basdevant, A. (2003). "[Natural history of obesity]." Bull Acad Natl Med **187**(7): 1343-52; discussion 1352-5.

Baxevanis, A. D. (2001). "The Molecular Biology Database Collection: an updated compilation of biological database resources." Nucleic Acids Res **29**(1): 1-10.

Belkin, N. J., P. Kantor, et al. (1995). "Combining the evidence of multiple query representations for information retrieval." Inf. Process. Manage. **31**(3): 431-448.

Ben-Hur, A., D. Horn, et al. (2002). "Support vector clustering." J. Mach. Learn. Res. **2**: 125-137.

Ben-Hur, A. and W. S. Noble (2005). "Kernel methods for predicting protein-protein interactions." Bioinformatics **21**(suppl_1): i38-46.

Benabdeslem, K. (2006). Descendant hierarchical support vector machine for multi-class problems. International Joint Conferences on Neural Networks. IJCNN 2006, Vancouver**:** 1470-1475.

Benson, D. A., I. Karsch-Mizrachi, et al. (2008). "GenBank." Nucl. Acids Res. **36**(suppl_1): D25-30.

Berg, A. H., T. P. Combs, et al. (2002). "ACRP30/adiponectin: an adipokine regulating glucose and lipid metabolism." Trends Endocrinol Metab **13**(2): 84-9.

Bhattacharjee, A., W. G. Richards, et al. (2001). "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses." Proc. Natl. Acad. Sci. USA **98**(24): 13790-13795.

Bhopal, R. S. (2002). Concepts of epidemiology : an integrated introduction to the ideas, theories, principles, and methods of epidemiology. Oxford ; New York, Oxford University Press.

Bhopal, R. S. (2008). Concepts of epidemiology : integrating the ideas, theories, principles, and methods of epidemiology. Oxford ; New York, Oxford University Press.

Bidus, M. A., J. I. Risinger, et al. (2006). "Prediction of lymph node metastasis in patients with endometrioid endometrial cancer using expression microarray." Clin Cancer Res **12**(1): 83-8.

Birkland, A. and G. Yona (2006). "BIOZON: a hub of heterogeneous biological data." Nucl. Acids Res. **34**(suppl_1): D235-242.

Bishop, C. (2006). Pattern Recognition and Machine Learning (Information Science and Statistics), Springer.

Bishop, C. M., M. Svensen, et al. (1998). "GTM: The generative topographic mapping." Neural Computation **10**(1): 215-234.

Borda, J. C. (1781). "Mémoire sur les élections au scrutin." Histoire de l'Académie royale des sciences(Imprimerie royale).

Bray, G. A. (1996). "HEALTH HAZARDS OF OBESITY." Endocrinology & Metabolism Clinics of North America **25**(4): 907-919.

Breiman, L. (1996). "Bagging predictors." Mach. Learn. **24**(2): 123-140.

Breiman, L. (2001). "Random forests." Machine Learning **45**: 5 - 32.

Breiman, L., J. Friedman, et al. (1984). Classification and Regression Trees, {Chapman & Hall/CRC}.

Breiman, L., J. H. Friedman, et al. (1983). Classification and Regression Trees. Belmont, CA, Wadsworth Publishing Company.

Burges, C. J. C. (1998). "A Tutorial on Support Vector Machines for Pattern Recognition." Data Mining and Knowledge Discovery **2**(2): 121-167.

Cancello, R., C. Henegar, et al. (2005). "Reduction of Macrophage Infiltration and Chemoattractant Gene Expression Changes in White Adipose Tissue of Morbidly Obese Subjects After Surgery-Induced Weight Loss." Diabetes **54**(8): 2277-2286.

Carbon, S., A. Ireland, et al. (2009). "AmiGO: online access to ontology and annotation data." Bioinformatics **25**(2): 288-9.

Cardoso, F. (2003). "Microarray technology and its effect on breast cancer (re)classification and prediction of outcome." Breast Cancer Res **5**(6): 303-4.

Charles, M.-A., E. Eschwege, et al. (2008). "Monitoring the Obesity Epidemic in France: The Obepi Surveys 1997-2006." Obesity **16**(9): 2182-2186.

Chuang, H.-Y., H. Liu, et al. (2004). Identifying Significant Genes from Microarray Data. Proceedings of the 4th IEEE Symposium on Bioinformatics and Bioengineering, IEEE Computer Society.

Chuang, H.-Y., H. Liu, et al. (2004). Combination Methods in Microarray Analysis.

Clancey, W. J. (1985). Heuristic classification, Stanford University.

Clark, R. D., A. Strizhev, et al. (2002). "Consensus scoring for ligand/protein interactions." Journal of Molecular Graphics and Modelling **20**(4): 281-295.

Clemen, R. T. (1989). "Combining forecasts: A review and annotated bibliography." International Journal of Forecasting **5**(4): 559-583.

Cochrane, G., R. Akhtar, et al. (2007). "Priorities for nucleotide trace, sequence and annotation data capture at the Ensembl Trace Archive and the EMBL Nucleotide Sequence Database." Nucl. Acids Res.: gkm1018.

Codeco, C. T. and F. C. Coelho (2006). "Trends in cholera epidemiology." PLoS Med **3**(1): e42.

Collins, A. (2006). "Obesity Statistics: Weight Statistics — Adults, Children, Obesity-Related Diseases." from http://www.annecollins.com/obesity/statistics-obesity.htm.

Collins, A. (2006). "Worldwide Obesity: Trends Global Obesity Trends, Globesity the Growing Epidemic of Chronic

Overweight ", from http://www.annecollins.com/obesity/worldwide-obesity.htm.

Condorcet (1785). Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix. Imprimerie royale. paris, Imprimerie royale.

Cristianini, N. and J. Shawe-Taylor (2000). An introduction to support vector machines and other kernel-based learning methods. Cambridge ; New York, Cambridge University Press.

Dasarathy, B. V. (2000). Sensor Fusion Architectures Algorithms and Applications IV, SPIE Optical Engineering Press.

Day, W. H. E. and H. Edelsbrunner (1984). "Efficient Algorithms for Agglomerative Hierarchical-Clustering Methods." Journal of Classification **1**(1): 7-24.

Deegalla, S. and H. Boström (2007). Classification of Microarrays with kNN: Comparison of Dimensionality Reduction Methods. Intelligent Data Engineering and Automated Learning - IDEAL 2007**:** 800-809.

Dettling, M. and P. Buhlmann (2004). "Finding predictive gene groups from microarray data." Journal of Multivariate Analysis **90**(1): 106-131.

Diaz-Uriarte, R. and S. Alvarez de Andres (2006). "Gene selection and classification of microarray data using random forest." BMC Bioinformatics **7**(1): 3.

Dietterich, T. G. (2000). "Ensemble methods in machine learning." Multiple Classifier Systems **1857**: 1-15.

Dudoit, S. and J. Fridlyand (2003). "Classification in microarray experiments." Statistical analysis of gene expression microarray data: 93 - 158.

Dudoit, S., J. Fridlyand, et al. (2002). "Comparison of discrimination methods for the classification of tumors suing gene expression data." J Am Stat Assoc **97**(457): 77 - 87.

Dudoit, S., J. Fridlyand, et al. (2002). "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data." Journal of the American Statistical Association **97**: 77-87.

Duerr, B., W. Haettich, et al. (1980). "A combination of statistical and syntactical pattern recognition applied to classification of unconstrained handwritten numerals." Pattern Recognition **12**(3): 189-199.

Duin, R. P. W. and D. M. J. Tax (2000). "Experiments with classifier combining rules." Multiple Classifier Systems **1857**: 16-29.

Dwork, C., R. Kumar, et al. (2001). Rank aggregation methods for the Web. Proceedings of the 10th international conference on World Wide Web. Hong Kong, Hong Kong, ACM.

Efron, B., and Robert J. Tibshirani (1993). An Introduction to the Bootstrap. New York, Chapman and Hall.

Egmont-Petersen, M., W. R. M. Dassen, et al. (1999). "Sequential selection of discrete features for neural networks - A Bayesian approach to building a cascade." Pattern Recognition Letters **20**(11-13): 1439-1448.

Ein-Dor, L., I. Kela, et al. (2005). "Outcome signature genes in breast cancer: is there a unique set?" Bioinformatics **21**(2): 171-178.

Ein-Dor, L., I. Kela, et al. (2005). "Outcome signature genes in breast cancer: is there a unique set?" Bioinformatics **21**: 171 - 178.

Eisen, M. B., P. T. Spellman, et al. (1998). "Cluster analysis and display of genome-wide expression patterns." Proceedings of the National Academy of Sciences of the United States of America **95**(25): 14863-14868.

Ellsworth, D. L. and T. A. Manolio (1999). "The emerging importance of genetics in epidemiologic research II. Issues in study design and gene mapping." Ann Epidemiol **9**(2): 75-90.

Ellsworth, D. L. and T. A. Manolio (1999). "The Emerging Importance of Genetics in Epidemiologic Research III. Bioinformatics and statistical genetic methods." Ann Epidemiol **9**(4): 207-24.

Ellsworth, D. L. and T. A. Manolio (1999). "The emerging importance of genetics in epidemiologic research. I. Basic concepts in human genetics and laboratory technology." Ann Epidemiol **9**(1): 1-16.

Fagin, R., R. Kumar, et al. (2003). Efficient similarity search and classification via rank aggregation. Proceedings of the 2003 ACM SIGMOD international conference on Management of data. San Diego, California, ACM.

Ferri, C., P. Flach, et al. (2004). Delegating classifiers. Proceedings of the twenty-first international conference on Machine learning. Banff, Alberta, Canada, ACM.

Ferri, C. and J. Hernandez-Orallo (2004). Cautious Classifiers. Proceedings of the 1st International Workshop on ROC Analysis in Artificial Intelligence, Valencia, Spain.

Fisher, R. A. (1936). "The use of multiple measurements in taxonomic problems." Annals of Eugenics **7**: 179-188.

Frank, E. and I. H. Witten (1998). Generating Accurate Rule Sets Without Global Optimization. Proceedings of the Fifteenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc.

Fraser, D. W. (1998). "Vitamins and vitriol: W.L. Braddon's epidemiology of Beriberi." Am J Epidemiol **148**(6): 519-27.

Freund, Y. and R. E. Schapire (1999). "A short introduction to boosting." Journal of Japanese Society for Artificial Intelligence, **14**(5): 771-780.

Friedel, C. C. (2005). On abstaining classifiers, Ludwig-Maximilians-Universitat. **Master's thesis**.

Friedel, C. C., R. Ulrich, et al. (2006). "Cost Curves for Abstaining Classifiers." Proceedings of the ICML 2006 workshop on ROC Analysis in Machine Learning Pittsburgh, PA.

Furey, T. S., N. Cristianini, et al. (2000). "Support vector machine classification and validation of cancer tissue samples using microarray expression data." Bioinformatics **16**(10): 906 - 914.

Galperin, M. Y. (2007). "The Molecular Biology Database Collection: 2007 update." Nucleic acids research **35**(Database issue): D3-4.

Gama, J. and P. Brazdil (2000). "Cascade generalization." Machine Learning **41**(3): 315-343.

Gevaert, O., S. Van Vooren, et al. (2008). "Integration of microarray and textual data improves the prognosis prediction of breast, lung and ovarian cancer patients." Pac Symp Biocomput: 279-90.

Gohlke, H. and G. Klebe (2001). "Statistical potentials and scoring functions applied to protein-ligand binding." Current Opinion in Structural Biology **11**(2): 231-235.

Goldfine, A. B. and C. R. Kahn (2003). "Adiponectin: linking the fat cell to insulin sensitivity." Lancet **362**(9394): 1431-2.

Golub, T. R., D. K. Slonim, et al. (1999). "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring." Science **286**(5439): 531-537.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring." Science **286**(5439): 531-537.

Guerre-Millo, M. (2002). "Adipose tissue hormones." J Endocrinol Invest **25**(10): 855-61.

Guyatt, G., S. Walter, et al. (1987). "Measuring change over time: assessing the usefulness of evaluative instruments." J Chronic Dis **40**(2): 171-8.

Hadzikadic, M., A. Hakenewerth, et al. (1996). "Concept formation vs. logistic regression: predicting death in trauma patients." Artificial Intelligence in Medicine **8**(5): 493-504.

Harima, Y., A. Togashi, et al. (2004). "Prediction of outcome of advanced cervical cancer to thermoradiotherapy according to expression profiles of 35 genes selected by cDNA microarray analysis." Int J Radiat Oncol Biol Phys **60**(1): 237-48.

Hart, R. O. D. a. P. E. (1973). Pattern Classication and Scene Analysis, John Wiley & Sons.

Hartigan, J. A. (1975). Clustering algorithms. New York,, Wiley.

Henegar, C., J. Tordjman, et al. (2008). "Adipose tissue transcriptomic signature highlights the pathological relevance of extracellular matrix in human obesity." Genome Biology **9**(1): R14.

Hermjakob, H., L. Montecchi-Palazzi, et al. (2004). "IntAct: an open source molecular interaction database." Nucl. Acids Res. **32**(suppl_1): D452-455.

Ho, J. W., M. Stefani, et al. (2008). "Differential variability analysis of gene expression and its application to human diseases." Bioinformatics **24**(13): i390-8.

Ho, T. K. (2002). Multiple Classifier Combination: Lessons and Next Steps, World Scientific.

Ho, T. K., J. J. Hull, et al. (1994). "Decision Combination in Multiple Classifier Systems." IEEE Trans. Pattern Anal. Mach. Intell. **16**(1): 66-75.

Hsu, D. F. and I. Taksa (2005). "Comparing Rank and Score Combination Methods for Data Fusion in Information Retrieval." Inf. Retr. **8**(3): 449-480.

Ibraev, U., P. Kantor, et al. (2001). "Counter-intuitive Cases of Data Fusion in Information Retrieval." Proceedings of the ASIST Annual Meeting **38**: 31-45.

Ibraev, U., K. B. Ng, et al. (2002). "Exploration of a Geometric Model of Data Fusion." Proceedings of the ASIST Annual Meeting **39**: 124-29.

Izmirlian, G. (2004). "Application of the random forest classification algorithm to a SELDI-TOF proteomics study in the setting of a cancer prevention trial." Ann NY Acad Sci **1020**: 154 - 174.

Juge-Aubry, C. E., E. Henrichot, et al. (2005). "Adipose tissue: a regulator of inflammation." Best Pract Res Clin Endocrinol Metab **19**(4): 547-66.

Kanehisa, M., M. Araki, et al. (2008). "KEGG for linking genomes to life and the environment." Nucl. Acids Res. **36**(suppl_1): D480-484.

Kerem, B., J. M. Rommens, et al. (1989). "Identification of the cystic fibrosis gene: genetic analysis." Science **245**(4922): 1073-80.
Khoury, M. J. (1997). "Genetic epidemiology and the future of disease prevention and public health." Epidemiol Rev **19**(1): 175-80.
Khoury, M. J. and Q. Yang (1998). "The future of genetic studies of complex human diseases: an epidemiologic perspective." Epidemiology **9**(3): 350-4.
Kim, H.-C., S. Pang, et al. (2003). "Constructing support vector machine ensemble." Pattern Recognition **36**(12): 2757-2767.
Kittler, J. and F. M. Alkoot (2003). "Sum Versus Vote Fusion in Multiple Classifier Systems." IEEE Trans. Pattern Anal. Mach. Intell. **25**(1): 110-115.
Kittler, J., M. Hatef, et al. (1998). "On combining classifiers." Ieee Transactions on Pattern Analysis and Machine Intelligence **20**(3): 226-239.
Klaus, S. and J. Keijer (2004). "Gene expression profiling of adipose tissue:: individual, depot-dependent, and sex-dependent variabilities." Nutrition **20**(1): 115-120.
Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. IJCAI.
Kohonen, T. (1995). Self-organizing maps. Berlin ; New York, Springer.
Kohonen, T. (1998). "The self-organizing map." Neurocomputing **21**(1-3): 1-6.
Koza, R. A., L. Nikonova, et al. (2006). "Changes in gene expression foreshadow diet-induced obesity in genetically identical mice." PLoS Genetics **2**(5): 769-780.
Kuncheva, L. I. (2002). "Switching between selection and fusion in combining classifiers: An experiment." Ieee Transactions on Systems Man and Cybernetics Part B-Cybernetics **32**(2): 146-156.
Kuncheva, L. I. (2002). "A theoretical study on six classifier fusion strategies." Ieee Transactions on Pattern Analysis and Machine Intelligence **24**(2): 281-286.
Kuncheva, L. I. (2004). Combining Pattern Classifiers: Methods and Algorithms, Wiley-Interscience.
Kuriakose, M., W. Chen, et al. (2004). "Selection and validation of differentially expressed genes in head and neck cancer." Cellular and Molecular Life Sciences CMLS **61**: 1372-1383.
Lanckriet, G. R. G. (2004). "Learning the Kernel Matrix with Semidefinite Programming." Journal of Machine Learning Research **5**: 27-72.
Larsen, J. K., R. Geenen, et al. (2004). "Personality as a Predictor of Weight Loss Maintenance after Surgery for Morbid Obesity." Obesity **12**(11): 1828-1834.
Lebbah, M., F. Badran, et al. (2000). Topological Map for Binary Data. ESANN Bruges**:** 267-272.
Lebbah, M., C. Chabanon, et al. (2002). Categorical Topological Map. Artificial Neural Networks — ICANN 2002**:** 793-794.
Lebbah, M., A. Chazottes, et al. (2005). Mixed Topological Map. ESANN**:** 357-362.
Lee, T. J., Y. Pouliot, et al. (2006). "BioWarehouse: a bioinformatics database warehouse toolkit." BMC Bioinformatics **7**: 170.
Lee, Y. C., W. J. Lee, et al. (2007). "Prediction of successful weight reduction after bariatric surgery by data mining technologies." Obes Surg **17**(9): 1235-41.
Lemoine, S., F. Combes, et al. (2006). "Goulphar: rapid access and expertise for standard two-color microarray normalization methods." BMC Bioinformatics **7**(1): 467.

Lewis, D. P., T. Jebara, et al. (2006). "Support vector machine learning from heterogeneous data: an empirical analysis using protein sequence and structure." Bioinformatics **22**(22): 2753-2760.

Liaw, A. and M. Wiener (2002). "Classification and regression by randomForest." Rnews **2**: 18 - 22.

Lin, S. M., J. Devakumar, et al. (2006). "Improved prediction of treatment response using microarrays and existing biological knowledge." Pharmacogenomics **7**(3): 495-501.

Liolios, K., K. Mavromatis, et al. (2008). "The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata." Nucl. Acids Res. **36**(suppl_1): D475-479.

Liu, R. and B. Yuan (2001). "Multiple classifiers combination by clustering and selection." Information Fusion **2**: 163-168(6).

Long, F. and C. Ding (2005). "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy." IEEE Trans. Pattern Anal. Mach. Intell. **27**(8): 1226-1238.

Long, P. M. and V. B. Vega (2003). "Boosting and microarray data." Machine Learning **52**(1-2): 31-44.

Loos, R. J. F. and T. Rankinen (2005). "Gene-Diet Interactions on Body Weight Changes." Journal of the American Dietetic Association **105**(5, Supplement 1): 29-34.

Ludmila I. Kuncheva (2004). Multiple Classifier Systems. Combining Pattern Classifiers**:** 101-110.

Lukasova, A. (1979). "Hierarchical Agglomerative Clustering Procedure." Pattern Recognition **11**(5-6): 365-381.

M. J. Moreno-Aliaga, J. L. S., A. Marti, J. A. Martيnez, (2005). "Does weight loss prognosis depend on genetic make-up?" Obesity Reviews **6**(2): 155-168.

Maggard, M. A., L. R. Shugarman, et al. (2005). "Meta-analysis: surgical treatment of obesity." Ann Intern Med **142**(7): 547-59.

Maglott, D., J. Ostell, et al. (2007). "Entrez Gene: gene-centered information at NCBI." Nucl. Acids Res. **35**(suppl_1): D26-31.

Mangasarian, O. L. (1994). "Nonlinear Programming."

Marengo, E. and R. Todeschini (1993). "Linear Discriminant Hierarchical-Clustering - a Modeling and Cross-Validatable Divisive Clustering Method." Chemometrics and Intelligent Laboratory Systems **19**(1): 43-51.

Melnik, O., Y. Vardi, et al. (2004). "Mixed group ranks: Preference and confidence in classifier combination." Ieee Transactions on Pattern Analysis and Machine Intelligence **26**(8): 973-981.

Mercer, T. (1909). "Functions of positive and negative type and their connection with the theory of integral equations." Transaction of London Philosophy Society **A**(209): 415-446.

Mizrachi, I. K. (2008). Managing Sequence Data. Bioinformatics**:** 3-27.

Morrow, T. (2007). "Gene expression microarray improves prediction of breast cancer outcomes." Manag Care **16**(8): 51-2.

Mount, D. (2004). Bioinformatics: Sequence and Genome Analysis, {Cold Spring Harbor Laboratory Press}.

Mudunuri, U., A. Che, et al. (2009). "bioDBnet: the biological database network." Bioinformatics **25**(4): 555-6.

Mutch, D. M. and K. Clément (2006). "Unraveling the Genetics of Human Obesity." PLoS Genetics **2**(12): e188.

N. Finer, D. H. R., C. L. Renz, A. C. Hewkin, (2006). "Prediction of response to sibutramine therapy in obese non-diabetic and diabetic patients." Diabetes, Obesity and Metabolism **8**(2): 206-213.

NCHS (2005). Health, United States 2005. National Center for Health Statistics. Hyattsville, MD.

Ortega, J., M. Koppel, et al. (2001). "Arbitrating Among Competing Classifiers Using Learned Referees." Knowledge and Information Systems **3**(4): 470-490.

Pang, S. N., D. J. Kim, et al. (2005). "Face membership authentication using SVM classification tree generated by membership-based LLE data partition." Ieee Transactions on Neural Networks **16**(2): 436-446.

Parkinson, H., M. Kapushesky, et al. (2007). "ArrayExpress--a public database of microarray experiments and gene expression profiles." Nucl. Acids Res. **35**(suppl_1): D747-750.

Patil, G. P. and C. Taillie (2004). "Multiple indicators, partially ordered sets, and linear extensions: Multi-criterion ranking and prioritization." Environmental and Ecological Statistics **11**: 199-228.

Pavlidis, P., J. Weston, et al. (2001). Gene functional classification from heterogeneous data Proceedings of the fifth annual international conference on Computational biology C Montreal, Quebec, Canada, ACM Press.

Pégorier, J. P. (2007). Le tissue adipeux: Son rôle dans les maladies métaboliques. Traité de nutrition artificielle de l'adulte**:** 341-352.

Perez-Diez, A., A. Morgun, et al. (2007). Microarrays for Cancer Diagnosis and Classification. Microarray Technology and Cancer Gene Profiling**:** 74-85.

Perrone, M. P. and L. N. Cooper (1993). When networks disagree: Ensemble methods for hybrid neural networks, Chapman and Hall.

Perusse, L. and C. Bouchard (2000). "Gene-diet interactions in obesity." Am J Clin Nutr **72**(5): 1285S-1290.

Petersen, M., M. A. Taylor, et al. (2005). "Randomized, multi-center trial of two hypo-energetic diets in obese subjects: high- versus low-fat content." Int J Obes **30**(3): 552-560.

Platt, J. (1999). Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. Advances in Kernel Methods - Support Vector Learning**.** B. Schölkopf, C. J. C. Burges and A. J. Smola, MIT Press**:** 185-208.

Plomin, R., M. J. Owen, et al. (1994). "The genetic basis of complex human behaviors." Science **264**(5166): 1733-9.

Pomeroy, S. L., P. Tamayo, et al. (2002). "Prediction of central nervous system embryonal tumour outcome based on gene expression." Nature **415**(6870): 436-442.

Prifti, E., J.-D. Zucker, et al. (2008). "FunNet: an integrative tool for exploring transcriptional interactions." Bioinformatics **24**(22): 2636-2638.

Raz, I., R. Eldor, et al. (2005). "Diabetes: insulin resistance and derangements in lipid metabolism. Cure through intervention in fat transport and storage." Diabetes Metab Res Rev **21**(1): 3-14.

Rimkus, C., J. Friederichs, et al. (2008). "Microarray-based prediction of tumor response to neoadjuvant radiochemotherapy of patients with locally advanced rectal cancer." Clin Gastroenterol Hepatol **6**(1): 53-61.

Ripley, B. D. and N. L. Hjort (1995). Pattern Recognition and Neural Networks, Cambridge University Press.

Risch, N. and K. Merikangas (1996). "The future of genetic studies of complex human diseases." Science **273**(5281): 1516-7.

Rissanen, A., M. Lean, et al. (2003). "Predictive value of early weight loss in obesity management with orlistat: an evidence-based assessment of prescribing guidelines." Int J Obes Relat Metab Disord **27**(1): 103-109.

Roli, F., S. Raudys, et al. (2002). "An experimental comparison of fixed and trained fusion rules for crisp classifier outputs." Multiple Classifier Systems **2364**: 232-241.

Schena, M., D. Shalon, et al. (1995). "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray." Science **270**(5235): 467-470.

Schèolkopf, B., C. J. C. Burges, et al. (1999). Advances in kernel methods support vector learning. Cambridge, Mass., MIT Press.

Scherer, P. E., S. Williams, et al. (1995). "A Novel Serum Protein Similar to C1q, Produced Exclusively in Adipocytes." J. Biol. Chem. **270**(45): 26746-26749.

Schölkopf, B., C. Burges, et al., Eds. (1998). Advances in Kernel Methods - Support Vector Machines. Cambridge, MA.

Scott, D. J., L. Villegas, et al. (2003). "Intraoperative ultrasound and prophylactic ursodiol for gallstone prevention following laparoscopic gastric bypass." Surg Endosc **17**(11): 1796-802.

Seewald, A. K., J. F\, et al. (2001). An Evaluation of Grading Classifiers. Proceedings of the 4th International Conference on Advances in Intelligent Data Analysis, Springer-Verlag.

Selvanayagam, Z. E., T. H. Cheung, et al. (2004). "Prediction of chemotherapeutic response in ovarian cancer with DNA microarray expression profiling." Cancer Genet Cytogenet **154**(1): 63-6.

Shah, S. P., Y. Huang, et al. (2005). "Atlas - a data warehouse for integrative bioinformatics." BMC Bioinformatics **6**: 34.

Shipp, M. A., K. N. Ross, et al. (2002). "Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning." Nat Med **8**(1): 68-74.

Signorini DF, A. P., Jones PA, Wardlaw JM, Miller JD (1999). "Predicting survival using simple clinical variables: a case study in traumatic brain injury." J Neurol Neurosurg Psychiatry **66**(1): 20-5.

Sjostrom, L., A. K. Lindroos, et al. (2004). "Lifestyle, diabetes, and cardiovascular risk factors 10 years after bariatric surgery." N Engl J Med **351**(26): 2683-93.

Smyth, G. K. and T. Speed (2003). "Normalization of cDNA microarray data." Methods **31**(4): 265-273.

Sorensen, T. I. A., P. Boutin, et al. (2006). "Genetic Polymorphisms and Weight Loss in Obesity: A Randomised Trial of Hypo-Energetic High- versus Low-Fat Diets." PLoS Clinical Trials **1**(2): e12.

Southern, E. M. (1975). "Detection of specific sequences among DNA fragments separated by gel electrophoresis." J Mol Biol **98**(3): 503-17.

Stolley, P. D. and T. Lasky (1995). Investigating disease patterns : the science of epidemiology. New York, Scientific American Library : Distributed by W.H. Freeman.

Sugawara, H., O. Ogasawara, et al. (2008). "DDBJ with new system and face." Nucl. Acids Res. **36**(suppl_1): D22-24.

Sungmoon Cheong, S. H. O. and S.-Y. Lee (2004). "Face membership authentication using SVM classification tree generated by membership-based LLE data partition." Neural Information Processing. Letters and Reviews **2**(3): 47-51.

Svetnik, V., A. Liaw, et al. (2004). "Application of Breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules." Multiple Classier Systems, Fifth International Workshop, MCS 2004, Proceedings, 9-11 June 2004, Cagliari, Italy. Lecture Notes in Computer Science, Springer **3077**: 334 - 343.

Tamayo, P., D. Slonim, et al. (1999). "Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation." Proceedings of the National Academy of Sciences of the United States of America **96**(6): 2907-2912.

Tax, D. M. J., M. van Breukelen, et al. (2000). "Combining multiple classifiers by averaging or by multiplying?" Pattern Recognition **33**(9): 1475-1485.

Team, R. D. C. (2003). R: A language and environment for statistical computing. Vienna, Austria, R Foundation for Statistical Computing.

The Gene Ontology Consortium (2008). "The Gene Ontology project in 2008." Nucl. Acids Res. **36**(suppl_1): D440-444.

The UniProt Consortium (2008). "The Universal Protein Resource (UniProt)." Nucl. Acids Res. **36**(suppl_1): D190-195.

Tibben, A., H. J. Duivenvoorden, et al. (1994). "Psychological effects of presymptomatic DNA testing for Huntington's disease in the Dutch program." Psychosom Med **56**(6): 526-32.

Tibshirani, R., T. Hastie, et al. (2002). "Diagnosis of multiple cancer types by shrunken centroids of gene expression." Proceedings of the National Academy of Sciences of the United States of America **99**(10): 6567-6572.

Trevor Hastie, R. T., and Jerome Friedman (2001). The Elements of Statistical Learning, Springer.

Trevor Hastie, S. R., Robert Tibshirani, Ji Zhu (2004). "The Entire Regularization Path for the Support Vector Machine." Journal of Machine Learning Research **5**: 1391--1415.

Troyanskaya, O., M. Cantor, et al. (2001). "Missing value estimation methods for DNA microarrays." Bioinformatics **17**(6): 520-525.

Tseng, Y. H., A. J. Butte, et al. (2005). "Prediction of preadipocyte differentiation by gene expression reveals role of insulin receptor substrates and necdin." Nature Cell Biology **7**(6): 601-U22.

Tubbs, J. D. and W. O. Alltop (1991). "Measures of Confidence Associated with Combining Classification Results." Ieee Transactions on Systems Man and Cybernetics **21**(3): 690-692.

Tumer, K. and J. Ghosh (1999). Linear and order statistics combiners for pattern classification. Combining Artificial Neural Nets, Springer-Verlag.

Tusher, V. G., R. Tibshirani, et al. (2001). "Significance analysis of microarrays applied to the ionizing radiation response." Proceedings of the National Academy of Sciences of the United States of America **98**(9): 5116-5121.

Vaessen, N. and C. M. van Duijn (2001). "Opportunities for population-based studies of complex genetic disorders after the human genome project." Epidemiology **12**(3): 360-4.

Vapnik, V. N. (1995). The Nature of Statistical Learning Theory. New York, Springer.
Vesanto, J. and E. Alhoniemi (2000). "Clustering of the self-organizing map." Ieee Transactions on Neural Networks **11**(3): 586-600.
Vesanto, J., J. Himberg, et al. (1999). Self-organizing map in Matlab: the SOM toolbox. In Proceedings of the Matlab DSP Conference.
Viguerie, N., C. Poitou, et al. (2005). "Transcriptomics applied to obesity and caloric restriction." Biochimie **87**(1): 117-123.
Viguerie, N., H. Vidal, et al. (2005). "Adipose tissue gene expression in obese subjects during low-fat and high-fat hypocaloric diets." Diabetologia **48**(1): 123-131.
Wahba, G., Y. Lin, et al. (2000). Generalized Approximate Cross Validation for Support Vector Machines, or, Another Way to Look at Margin-Like Quantities. Advances in Large Margin Classifiers. A. Smola, P. Bartlett, B. Schölkopf and P. Schuurmans, MIT Press**:** 297-309.
Wang, J., J. Delabie, et al. (2002). "Clustering of the SOM easily reveals distinct gene expression patterns: results of a reanalysis of lymphoma study." BMC Bioinformatics **3**(1): 36.
Watanabe, T., T. Kobunai, et al. (2007). "Gene expression signature and the prediction of ulcerative colitis-associated colorectal cancer by DNA microarray." Clin Cancer Res **13**(2 Pt 1): 415-20.
Watanabe, T., Y. Komuro, et al. (2006). "Prediction of sensitivity of rectal cancer cells in response to preoperative radiotherapy by DNA microarray analysis of gene expression profiles." Cancer Res **66**(7): 3370-4.
Webb, A. R. (2002). Statistical pattern recognition. West Sussex, England ; New Jersey, Wiley.
Wei, W., L. Xin, et al. (2004). "A Hybrid SOM-SVM Method for Analyzing Zebra Fish Gene Expression." icpr **02**: 323-326.
WHO (2006). World Health Statistics. W. H. Organization. Geneva**:** pp. 42-48.
Wolpert, D. H. (1992). "Stacked Generalization." Neural Networks **5**(2): 241-259.
Wolpert, D. H. (1996). "The existence of A priori distinctions between learning algorithms." Neural Computation **8**(7): 1391-1420.
Wolpert, D. H. (1996). "The lack of A priori distinctions between learning algorithms." Neural Computation **8**(7): 1341-1390.
Wong, Y. F., Z. E. Selvanayagam, et al. (2003). "Expression genomics of cervical cancer: molecular classification and prediction of radiotherapy response by DNA microarray." Clin Cancer Res **9**(15): 5486-92.
Woodward, M. (1999). Epidemiology : study design and data analysis. Boca Raton, FL, Chapman & Hall/CRC Press.
Wu, T., J. Wang, et al. (2006). "NPInter: the noncoding RNAs and protein related biomacromolecules interaction database." Nucl. Acids Res. **34**(suppl_1): D150-152.
Xenarios, I., L. Salwinski, et al. (2002). "DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions." Nucl. Acids Res. **30**(1): 303-305.
Xing, E. P., M. I. Jordan, et al. (2001). Feature selection for high-dimensional genomic microarray data. Proceedings of the Eighteenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc.
Xu, L., A. Krzyzak, et al. (1992). "Methods of Combining Multiple Classifiers and Their Applications to Handwriting Recognition." IEEE Trans. Systems Man Cybern **22**(3): 418-435.

Yacoub, M., F. Badran, et al. (2001). "A topological hierarchical clustering: Application to ocean color classification." Artificial Neural Networks-Icann 2001, Proceedings **2130**: 492-499.

Yager, R. R. (2000). "Intelligent control of the hierarchical agglomerative clustering process." Ieee Transactions on Systems Man and Cybernetics Part B-Cybernetics **30**(6): 835-845.

Yang, J.-M., Y.-F. Chen, et al. (2005). "Consensus Scoring Criteria for Improving Enrichment in Virtual Screening." J. Chem. Inf. Model. **45**(4): 1134-1146.

Yang, Y. H., S. Dudoit, et al. (2002). "Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation." Nucl. Acids Res. **30**(4): e15-.

Zhang, Y., R. Proenca, et al. (1994). "Positional cloning of the mouse obese gene and its human homologue." Nature **372**(6505): 425-432.